The Geometry of Colour

Lucas Wilkins

Submitted for the degree of Doctor of Philosophy

University of Sussex

April 2012

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another university for the award of any other degree.

Lucas Wilkins (April 28, 2013)

Abstract

This thesis explores the geometric description of animal colour vision. It examines the relationship of colour spaces to behavior and to physiology. I provide a derivation of, and explore the limits of, geometric spaces derived from the notion of risk and uncertainty aversion as well as the geometric objects that enumerate the variety of achievable colours. Using these principles I go on to explore evolutionary questions concerning colourfulness, such as aposematism, mimicry and the idea of aesthetic preference.

Acknowledgements

I would like to thank my supervisors Danny Osorio and Thomas Nowotny for their continued support. I would like to thank Danny firstly for giving me a doctoral position, which by any account, is nothing short of a privilege. Secondly, I would like to thank him for his honest advice and always allowing me the freedom to investigate in the directions I choose. I would like to thank Thomas for his openness and rigor, and for many hours of whiteboard time.

I would like to thank all those in the Centre for Computational Neuroscience and Robotics for being essential parts such a diverse and interesting group. Of my colleagues in the CCNR, I would especially like to thank Greg Studer, Matt Egbert, Oliver Winks, Paul Graham, Jose Fernandez-Léon, Nathaniel Virgo, James Thorniley and Greg Corcoran for countless interesting and illuminating discussions. Further afield, I would like to thank Mike Beaton for helping to me understand minds less badly, and Johnathan Green for responding to some of my more abstruse questions concerning biology.

I would also like to thank Tim Nott and Tom Collett for their more diffuse influence. Both of whom remind me, in their own way, that there is more to science than data and logic.

I would like to thank my parents, Wendy and Bob Wilkins, for their continued love and support. I would particularly like to thank my father for raising me with a passion for science and putting me on a path which is always fascinating.

I would like to thank those who have kept me sane during the last few weeks of writing: Mike Beaton for abiding an unceasing flux of books, washing up and vegetables as well as my unsociable hours, and Jean Carrol for 'feeding me with hot food and kicking my arse'.

A number of people checked my writing for typographical errors - Danny, Tom, Mike and James - their bravery in the face of my punctuation shall not be forgotten.

I would like to thank the BBSRC for their financial support.

Contents

Technical Terms

Throughout this thesis I have tried to avoid specialist technical terms as much as possible, and to define them when I do use them. However, sometimes the definitions lie a long way from a given usage. It is also fairly difficult to avoid technical terms in mathematics. The following is a table of terms which I feel might aid the reading of this document. It is not complete, but I hope that it helps.

Term	Definition
Geometrising Transform	(coined) A transformation of a divergence to make it obey geometric axioms
n-chromat	(coined) An organism with n functioning photoreceptor cell classes i.e. the number cone cell types in daylight or rod cell types at night.
Colourfulness	(re-appropriated) A property of a colour similar to its saturation but defined in terms of the colour solid. This is different to the existing formal interpretation of the term. Colour science has formal meanings for all terms that might well describe this dimension making it necessary to 'overload' some or other term. This one is chosen as the existing meaning is least relevant for this work.
Monochromat	An organism that has only one functioning photoreceptor cell class.
Dichromat	An organism with two functioning photoreceptor cell classes.
Achromatic Vector	A vector that points in the direction from black to white
Metamer	Something that has the same colour as something else but a different visible spectrum.
Quantum Catch	The number or rate of photoisomerisation events within a photoreceptor cell.
Aposematism	Colours whose purpose is to warn an organisms predators of its toxicity.
Structural Colour	Colours formed by the microscopic structure of a surface, not directly by absorbance of light.

Preface

This thesis is about the geometry of colour, how it arises, what it means and what its consequences are. Specifically, it concerns an animal's colour vision as a factor in its behaviour and the evolution of their species.

Geometry is a tool for intuiting abstract concepts and structures.[1] But it is not merely a tool for providing intuition; it is also a tool for description which provides a methodology that can be used to provide concrete results. It is a practical and quantitative narrative.

This thesis is split into three parts. In the first part I look at traditional ideas of colour. First focusing on the history of colour theory then moving on to a generalisation of some well established concepts in colour theory. In the second part I examine the idea of perceptual spaces - of which colour spaces are a particular example. I develop a behaviourally justified information theoretic approach to colour space and use it to both examine existing spaces and derive my own. In the last part I consider evolutionary scenarios, firstly I ask why the signals found in the animal kingdom should be strongly coloured, and secondly, I enquire into the effect of structural colours on the evolution of warning signals.

Along with the geometric theme that runs though this thesis, there is also a focus on judgement. Here I take the attitude that this is something we are forced to consider when we ask questions concerning the behaviour of animals, and, that considering the basis of our judgements allows us to better understand what it is that we are discussing, illuminating the foundations upon which our theories stand.

[1] In the present case, the structures defined by information theory and physics.

P.1 Colour and Geometry

The first part of this thesis focuses on establishing the basic ideas of colour theory and on the large scale structure of colour spaces.

P.1.1 A Brief History of Colour Space

In this first chapter I outline the history of colour theory. I use this chapter to introduce the concepts of colour spaces and psychophysics.

P.1.2 The Shape of Colour Spaces

Much of what people consider to be the defining properties of colour spaces is their shape. The overall shape of a colour space is defined by the overlap of photoreceptor responses (making some signals impossible) and by the constraints of the class of spectra that we choose to consider. This chapter explores the properties of this shape.

P.2 Colour and Measurement

In the second major part I take the notion of behaviour as a goal directed process as the basis of a formal derivation of models describing experiments in colour. This part is focused on behaviour, neurophysiology, discrimination and judgement.

P.2.1 Measuring Living Things

In this chapter I aim to elucidate the meaning of information measures in terms of goal directed action. I provide a basis upon which we can ground beliefs, goals and actions in physical measurements. Eventually, this allows us to examine existing information theoretic quantities and decide upon their relevance to the understanding of behaviour and perception.

P.2.2 Geometry from Information

Beginning with the notion of a binary risk as a quantified representation of an agents goal I derive the general form of colour distances to which any particular colour based choice depends. These represent optimal[2] decision making processes under uncertainty. The resulting distances are information divergences and have some general properties that

[2]The optimality does not mean that they cannot be representative of heuristic processes. Indeed, in some cases they definitely are.

distinguish them from distances as usually conceived, this makes them a much broader class of measurement. I outline the various sub classes of these distances and note that when it is definable they all induce the same local metric on the colour space up to a multiplicative constant - this metric is the Fisher metric. From this we can see that the same metric occurs in all behavioural experiments for which the goal can be expressed as a binary risk. However, where finite colour distances are desired there is no uniquely defined measure of distance. But, even without knowing this, something can be said about what properties these distances should have and new behavioural predictions about generalisations can be made.

P.2.3 Local Models of Colour Vision

This chapter focuses on the correspondence between the model I propose in the previous section and those that currently exist. In a number of cases maximum entropy distributions can be used to show the correspondence between the theory and existing models (such as Chittka, 1992; Maxwell, 1860; Schrödinger, 1920; Vorobyev and Osorio, 1998). I also examine a case where uncertainty is removed (a deterministic limit) showing that the topology of the space induced by the metric is not what we would require from a colour space that matches either our notions of colour or physical plausibility. I claim that non-statistical spaces such as the Honeybee-Hexagon space of Chittka (1992) either make implicit assumptions or are otherwise unfounded. To remedy this I calculate the metric corresponding to a saturating photoreceptor from basic statistical principles. To finish with, I examine some general properties of the models discussed and observe that all have minimum discrimination information around the achromatic point. I also classify colour spaces by their ability to saturate, noting that most existing models do not have this property and that of those who do, excepting the model I provide, none are based on physical properties.

P.3 Colour in Nature

The last part of this thesis is about colour as a factor in evolution, particularly in animal communication.

P.3.1 The Colourfulness of Signals in Animal Communication

In this chapter I discuss the reasons why animal signals should be colourful. I argue that colourfulness should be valued in a similar way by all organisms on the basis of its ability

to inform. I argue that the degree of colourfulness allows us to judge the affordance of a given object. I present a number of cases where this is important, and argue that this kind of judgement – one that is made by all organisms in the same way – is a reason why we should think that Darwin is justified in using the word beauty in his works.

P.3.2 Structural Colours in Batesian Mimicry

The physical basis of structural colours is rather different to that of pigments. Humans observe this difference as the physical parameters affecting only the hue of a colour. This property serves as a source of non-linearity in evolutionary processes. Here, I visualise this in a simple model of structural colours and discuss the implications for mimicry and the formation of waring colours.

Part I

Colour and Geometry

Chapter 1

A Brief History of Colour Space

"Theres logic of colour, damn it all! The painter owes allegiance to that alone. Never to the logic of the brain."

Paul Cézanne
(Cézanne: a memoir with conversations, Joachim Gasquet)

1.1 Aristotle

Probably the most general definition of a colour space is "a geometric representation of colour". In this sense, colour spaces have a long history, dating back at least as far as ancient Greece. Aristotle provides us with one of the earliest examples by describing an ordering relationship for colours:[1]

> [...] colour has specific differences: therefore blackening and whitening differ specifically; but at all events every whitening will be specifically the same with every other whitening and every blackening with every other blackening. But white is not further subdivided by specific differences: hence any whitening is specifically one with any other whitening. Where it happens that the genus is at the same time a species, it is clear that the motion will then in a sense be one specifically though not in an unqualified sense [...]

Aristotle
Physics (∼350BC)
Book V, Part IV
See: Hardie and Gaye (1994)

[1] Although an ordering may be considered pre-geometric, he then goes on to associate colours with rational and irrational numbers.

He considered there to be a scale of colours, with a unique white at one end and one unique black at the other. These being related by processes which have been translated as "blackening" and "whitening". The exact nature of these processes were a matter of debate during his time, as was the nature of colour itself.

The notion of a colour space was not explicitly discussed by Aristotle, but he did approach the subject of colour in a manner which allows one to be defined.

> This being the true nature of mixture, it is plain that when bodies are mixed their colours also are necessarily mixed at the same time.
>
> *Aristotle*
> *On Sense and Sensible Objects (∼350BC)*
> *Part 4*
> *See: Beare (1994)*

Here he mentions what we now may consider to be the defining property of colour spaces (and psychophysical spaces in general) - a relationship between something physical and something perceived. Aristotle uses the physical processes of mixing to provide a process that changes objects in some quantifiable way[2] and then states that colours change in way which reflects this. The exploration of a physical parameter with an apparent quantity is explicit, therefore it seems reasonable to interpret Aristotle's "On Sense and Sensible Objects" as an investigation into the nature and consequence of this relationship, making him a very early pioneer of colour spaces.

1.2 The Renaissance

From the end of Classical Greece[3] to the beginning of the modern period there are few known descriptions of colour spaces.[4] Though there are exceptions, such as the space of

[2]Mixing is thoroughly explored by Aristotle. In his philosophy, in which all things are constituted by a finite number of atoms, mixable things may be mixed in integer ratios of their constituent parts. Degrees of mixing may be quantified by a rational number between zero and one. This approach to mixing is evident throughout "On Sense and Sensible Objects" (Beare, 1994). It is not unreasonable then, given that he describes elsewhere that certain colours correspond to irrational numbers, that we could construct a colour space with axes corresponding to the rational contributions of each fundamental (irrational) colour i.e. a space isomorphic to $\{1...n_i\}^m : \sum_i n_i = N$ for some number of unique colours m. This is contrary to the popular belief that he thought colours to be a 'product' of black and white.

[3]Alexander the Great, whose death is often taken to be the end of the Greek Classical period, died in 323 BC and Aristotle in 322 BC.

[4]Whether this is due to a general bias in scholarship towards the classical and modern periods, or genuinely reflects the evolution of colour theory is for someone more versed in history to decide.

Robert Grosseteste (ca. 1170-1253) which has recently drawn attention (Smithson et al., 2012).

Although geometry as a tool for understanding nature had existed since early Classical Greece, in the Renaissance it found a new role in perspective theory. Using geometrical theory the objective laws of the three dimensional space that people inhabit and can measure were transformed into egocentric laws on a two dimensional plane.[5] The laws of perspective found a central role in the painting of the time (Ackerman, 1980) and formed the basis for (at least one) architectural theory (Argan and Robb, 1946; Wittkower, 1953). It is probably not unfair to cite perspective as the oldest example of a geometric and psychophysical law.

It was in the mid 15$^{\text{th}}$ century Leon Battista Alberti published his well known colour system.[6][7] This system used four highly saturated 'primaries' whose combination was desaturated either with black or white depending on whether the current portion of a paintings subject was in light or shade (Bomford, 1995). The dislike that the Quattrocentos[8] had of dull painting lead Alberti to design this system in such a way to optimise the vividness of colour: "we all by nature love things that are open and bright; so we must the more firmly block the way in which it is easier to go wrong" (Ackerman, 1980). White and black were known to make colours less vivid, so they were added at the latest possible point. This leads us to naturally identify the first mixing (before addition of black or white) with the plane that in modern colour spaces that we would now describe as containing the colours of highest saturation. This system allows identification with features of many of the reflectance based colour spaces that are used today: a (hyper) plane consisting of the colours formed by the optimally saturated primaries provides a base for a bipyramid with the unique black and white points lying at the tips - the colours desaturating, and their variety decreasing, the nearer one moves to them.

The system of Alberti was still procedural, retaining in many ways the 'blackening' and 'whitening' of Aristotelian tradition (as Aristotle was the only work available at the

[5]It has been theorised that the subjective laws of perspective were favoured as a basis for architecture and painting, as it was thought that æsthetic value is bought about by the subject and that 'simulating' the subject, or at least, drafting or painting in a way that reflects qualities of subjective experience, results in a work of greater value (Argan and Robb, 1946).

[6]I use 'system' in the original sense of a method by which one adds coloured paint to a canvas, as opposed to the more scientific interpretation as a systematic arrangement of colour.

[7]He was, of course, not the only one to publish a colour system, but for the sake of brevity it will be the only system I discuss here.

[8]15$^{\text{th}}$ century Italian artists.

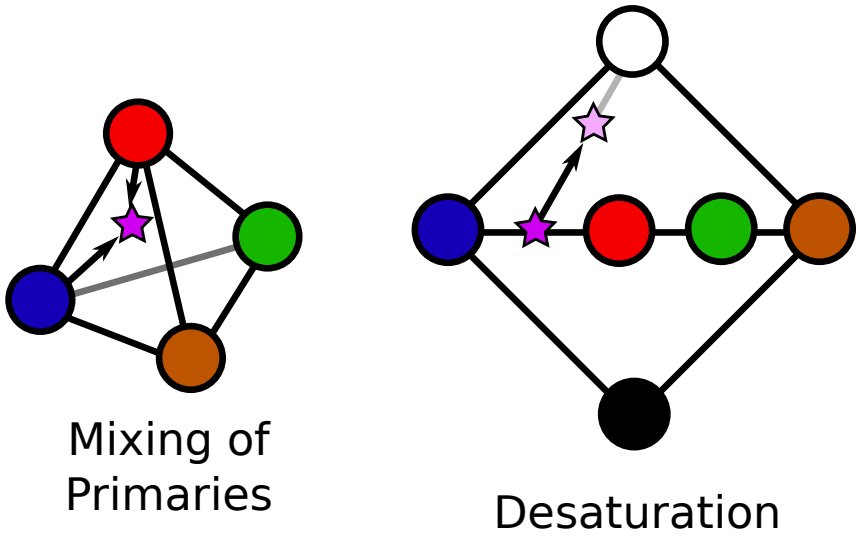

Figure 1.1: The colour system of Alberti. This shows the mixing of a lilac type colour (the star). First, the purest possible colour was mixed from four primaries, represented here in a tetrahedron. This colour is then desaturated by mixing black or white (white in this example). The tetrahedron from the first mixing is shown projected onto a line bisecting pure white and pure black with the body of the bipyramid containing all mixable colours being projected onto a diamond.

time Ackerman, 1980). It is not until after the Renaissance that we begin to see explicitly geometric colour spaces. However, it is good to remember that it was this period that saw the first use of geometry as a means to understand visual perception as well as the accurate documentation of gross features of colour spaces (in the form of colour systems).

1.3 The Enlightenment and Industrial Revolution

1.3.1 Newton and Goethe

Issac Newton is well known for his studies on light and a large part of his work was the study of the colour formed by thin films and prisms. Newton described light as a continuum of vibrations, using his study of the phenomena of thin film interference and dispersion to describe the visible part of the electromagnetic spectrum (Newton, 1665).

This physical understanding of light proves an indispensable tool for the description of colour. One can even consider the representation of the electromagnetic spectrum in figure 1.2 to be a colour space of kinds - in the sense that it maps a physical parameter to colour. Of course, this space does not pretend to represent the entirety of the colours which can be seen. Any light reflected from or transmitted through an object has a spectral composition

(an intensity value at every wavelength) which is determined by its physical properties and the spectral composition of the light incident upon it.

Dispersion, the phenomenon that Newton is most well known for investigating, provides a way of spatially separating wavelengths of light so that the intensity of each wavelength can be seen in isolation, the wavelength dependent nature of the intensity can then be observed spatially, for example in the thin dark (absorbance) bands seen in a sufficiently resolved spectrum of solar radiation. Colour is represented in space.

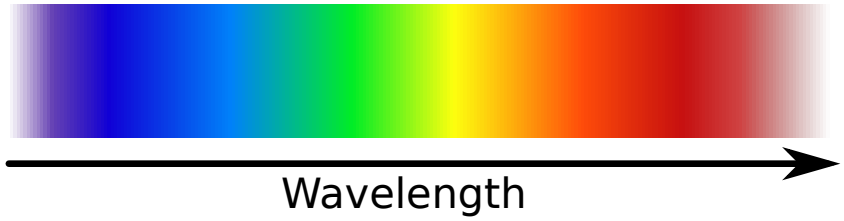

Figure 1.2: The visible spectrum. This representation can be seen as a mapping from the physical parameter (wavelength) to colours.

In a story about colour one cannot talk about the work of Newton without discussing the work of his 'opponent' Johann Wolfgang Von Goethe (although Goethe lived much later). Whereas Newton's understanding was physical and (arguably) reductionist, Goethe's understanding was at the phenomenological level; for example, where Newton considered darkness to be simply an absence of light, Goethe considered darkness to be a phenomenon of equal stature to - and with the same primacy as - lightness (Duck, 1988), see figures 1.3 and 1.4. Whilst this 'reification' of darkness was a major point of contention, what really separated the two was their modes of explanation. Goethe described many qualities of colour where on its own the Newtonian understanding of light was insufficient and a more holistic approach must be taken (such as phosphenes and the reddening of the visual field in hypoxic conditions; Theory of Colours, Eastlake trans. 1967). In many ways, the two were interested in very different things.

Unlike in the Renaissance, where the scientific understanding of perspective provided a harmony to the relationship between nature and personal experience (Argan and Robb, 1946), the conflict between the ideas of Newton and Goethe reflects as dissonance between the physical and the mental. Goethe's attack on Newton's ideas stemmed from his awareness of the phenomenological complexity of colour - a complexity that we are far from explaining even today. Nonetheless it is Newton's work, not Goethe's, that forms the basis of the modern scientific understanding of colour. The work of Goethe simply

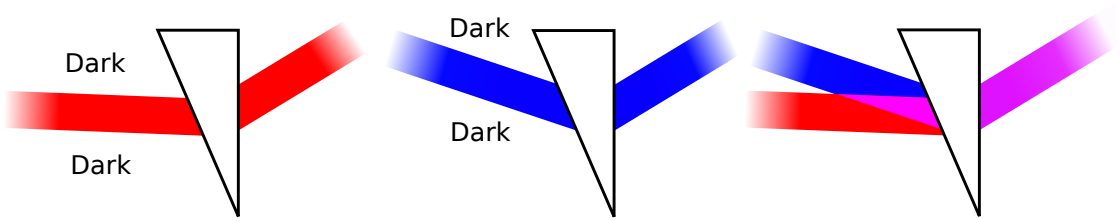

Figure 1.3: A representation of Goethe's argument against Newton's theory of colour. Here I have represented the dispersion of three light sources illuminating a prism from the right: short wavelength (blue), long wavelength (red) and a mixture of the two (purple). In the case of the mixture, pure red and pure blue are only seen in the areas which are not illuminated by the other wavelength. When considering a white light source, such as daylight, this has the consequence that colour is only observed in the 'dark fringes' of the transmitted beam. It is because of this that colour can be seen to occur where light and dark interact, see figure 1.4.

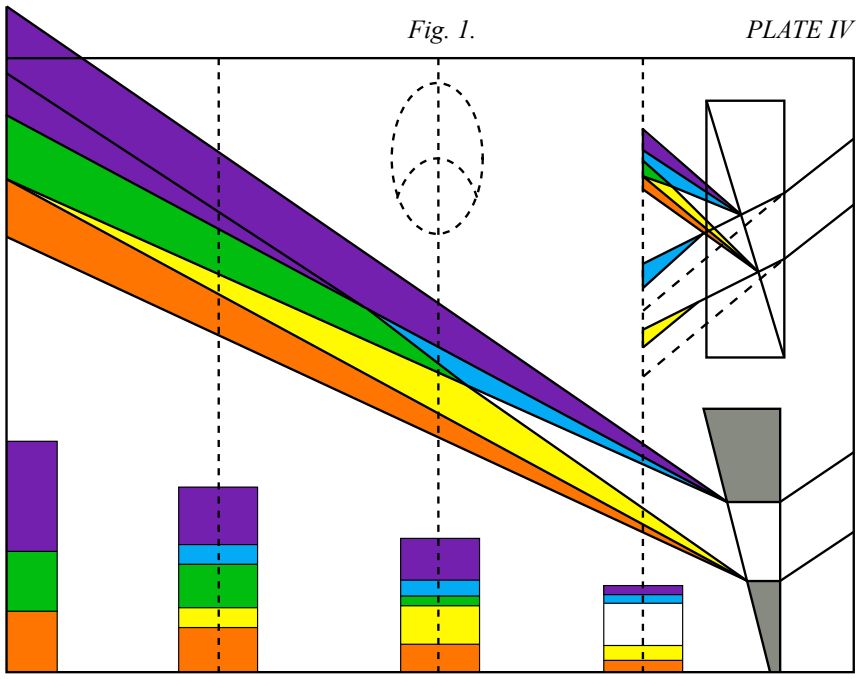

Figure 1.4: Figure from Theory of Colours (Eastlake, 1967) showing how colour occurs in the dark fringes of a transmitted beam. This is similar figure 1.3 but applied refraction of sunlight. [public domain image]

stands as a reminder that our scientific understanding of colour accounts for only a small part of our experience of it.

During the 17th and 18th there were a number of geometric conceptualisations of colour (see Shmid, 1948, for a concise account of the spaces of this period). Both Newton and

Goethe created colour spaces where hues were arranged in a circle, and a number of other geometric orderings of colours arose at this time. All of these were orderings of colour by empirical or procedural[9] similarity, rather than being based specifically on physiological knowledge. It is not until the development of trichromacy do we begin to see any significant physiological contributions to the geometry of colour spaces.

1.3.2 Trichromacy

In the 19th century, Thomas Young, and later Hermann von Helmholtz developed what is now known as the Young-Helmholtz theory of trichromacy. In this theory a colour is described by a triple of numbers determined by the relative intensity of light in different parts of the visual spectrum. After discussing Newton's theory of resonating particles in the retina[10], Young explained colour as the resonance of three types of particles with greater or lesser tendency to resonate with light of particular frequencies:

> [As] it is almost impossible to conceive each sensitive point of the retina to contain an infinite number of particles, each capable of vibrating in perfect unison with every possible undulation, it becomes necessary to consider the number limited, for instance, to three principal colours, red, yellow and blue, of which the number of undulations are related in magnitude nearly as the numbers 8,7,6; and that each of the particles is capable of being put in motion less or more forcibly [...]
>
> The Bakerian Lecture: *On the Theory of Light and Colours*
>
> *See: Young (1802)*

von Helmholtz (1896) took this idea further and provided a description of the sensitivity of each class of particle (now known to be opsin molecules) as a function of wavelength. The currently favoured choice of these functions are shown in figure 1.5. These functions correspond to the classes of cone cells within the human retina each of which contain only a single type of opsin. The degree to which each cell is activated for a given spectrum is

[9] Such as the process of pigment mixing used in the tetrahedral space (Farben-Pyramide) of Lambert (1772).

[10] My terminology. Newton originally conceived of rays of light as inducing superluminal vibrations in the æther. He then abandoned the vibrations of the æther in favour of vibrations of the medium in which they were travelling, i.e. light caused vibrations in whatever it was passing through. Newtons final understanding of light, however, was that instead of the vibration of the medium that it - or the medium in which it was travelling in - fell into 'fits' (periodic tendencies named after recurring fevers - *paroxysma* - of certain diseases especially malaria) which determined the ease of reflection or transmission (Shapiro, 1993, chap 3).

given by the summation of the intensity of light at each wavelength after multiplication by the sensitivity of each cell type to that wavelength. Formally, we write this as:

$$q^i = \int_\Lambda w^i(\lambda)s(\lambda)\mathrm{d}\lambda \tag{1.3.1}$$

where i indexes the cone type (for humans long, medium or short). The triple of numbers (vector) q is known as the quantum catch and can be thought of as a physical description of the colour. w is a weighting function (cone fundamental, see figure 1.5), s is the spectral composition (or just "spectrum": a positive real number at each value of λ). λ is a particular wavelength and Λ is the set of all wavelengths where the functions w and s are defined. This can also be written in terms of wavenumber, or frequency by change of the measure $\mathrm{d}\lambda$. Generally, it is expedient to take w^i to be dimensionless and s to be in units of photons per time per wavelength. With a vector representing colour, we have the necessary starting point for a physiologically based geometric account of colour.

Equation 1.3.1 is fundamental to any relationship between physical parameters and colour. It is minimal. It relates the most physical thing that one might reasonably accept as colour (the quantum catches) and the most colour like thing that one might reasonably accept as purely physical (spectra)[11].

Towards the end of the 19$^{\text{th}}$ century we have a theory of colour which is defined in terms of the physical and physiological understanding of the day. With this, we see the continuation of a trend that began with Newton. As the physical basis of colour became more and more understood, theories of colour became theories of physics, and the representations of colours became the representations favoured by physicists.

The colour space of James Clerk Maxwell sums up the progress made by 19$^{\text{th}}$ century understanding of colour. He maps colour to plane by normalising quantum catches by the sum of quantum catches (Maxwell, 1860), visualised in 1.6:

$$c^i = \frac{q^i}{\sum\limits_i q^i} \tag{1.3.2}$$

This projects all colours onto a triangle, which can then be viewed in two dimensions.

In the colour space of Maxwell the spectrum (which can be thought of as a one-dimensional colour space) is a one dimensional locus - known as the spectral line, or, monochromatic locus. This defines a boundary in which all colours lie[12].

[11] Of course, the physics/colour dichotomy is not well defined, I use these terms more descriptively than with any rigour. If pushed I would have to admit that this is not, in the sense described, as minimal as

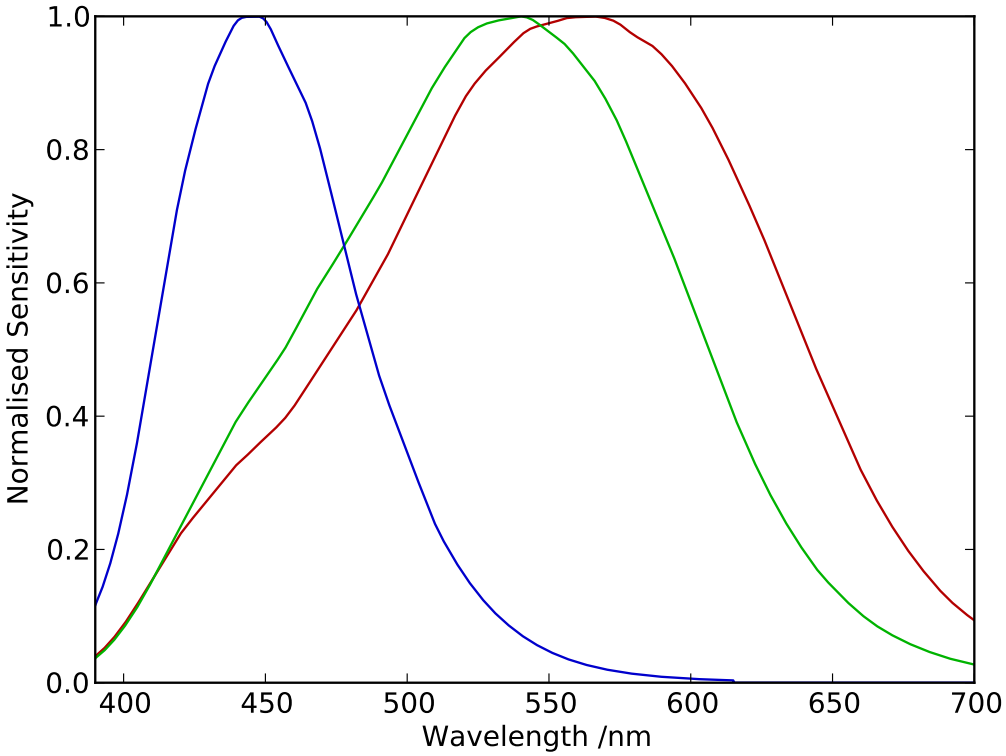

Figure 1.5: Modern functions used to determine the long wavelength (red), mid wavelength (green) and short wavelength (blue) coordinates of certain human colour spaces. These curves are psychophysically determined and thus do no correspond directly to quantum catch other than as a rough approximation. Shown here are the 2° cone fundamentals of Stockman and Sharpe (2000) in units of photons per wavelength.

1.3.3 The Rise of Psychophysical Theory

Around the same time as Helmholtz, Ernst Heinrich Weber and Gustav Fechner were starting a new branch of psychology: psychophysics. Their goal of was to measure perceived difference as it relates to physical difference. The most well known results are Weber's and Fechner's laws[13], which have been shown to occur in multiple modalities and tasks. The general form of both their laws, the Weber-Fechner law has been implicated in everything

one might hope.

[12] All spectra can be thought of as convex combinations of monochromatic lights. As all quantum catches are linear projections of spectra, colours are convex combinations of the colours corresponding to monochromatic lights. Before projection, all colours lie within a "cone" with a boundary of monochromatic lights at various intensities. When further projected using a perspective transform, they are found to lie within the the convex hull of the monochromatic locus (spectral line). See chapter 2 for a derivation

[13] Often just called the Weber-Fechner law

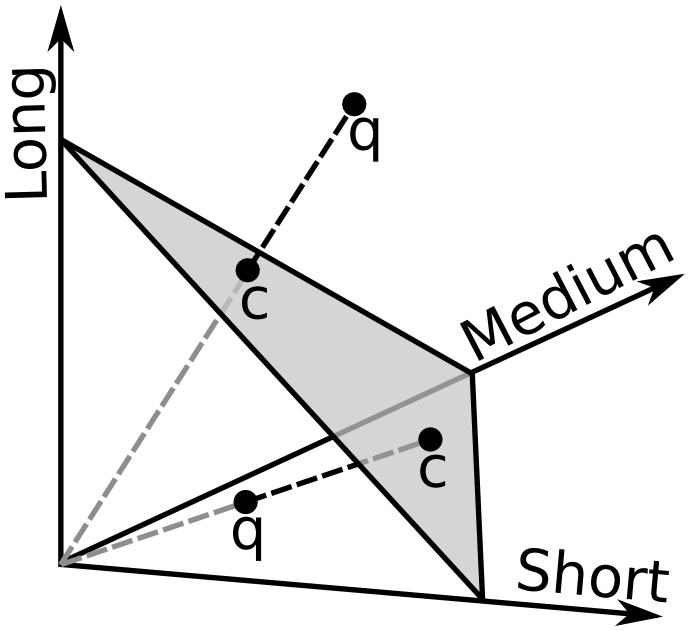

Figure 1.6: Illustration of the formation of Maxwell's triangle. The normalisation by the sum of quantum catches (q) moves them onto a (subset of a) triangle with corners at $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$. The points in this triangle corresponding to a particular q are labelled c. They are also given by the intersection of the triangle with a line from q to the origin.

from numerical cognition (Dehaene, 2007) to cell signalling (Cope, 1976). It has even been suggested that it is a general property of neurons[14](Deco and Rolls, 2006). One can state Weber's law as "the ability to discriminate stimuli is proportional to their magnitude" and Fechner's law as "the perceptual distance between two stimuli is proportional to their magnitude". They both admit the same formal representation, differing only in the assertions made about the domain of its applicability. A common form of the Weber and Fechner laws is:

$$\Delta x = kx \tag{1.3.3}$$

Where Δx is a physical difference between any equally discriminable (or perceptually different) quantities with (mean) magnitude x and where k is a constant of proportionality. Although both Weber's and Fechner's laws are very similar, they are none the less very different in their application, one is about discrimination - the ability to tell two things apart - the other is about perceptual distance - a *judgement* of difference. Discrimination can be used to great effect and its description with Weber's law is rather robust. Contrary to this, Fechner's law is more difficult to find or apply. In the theory of colour, there has

[14]There is obviously no reason to believe this.

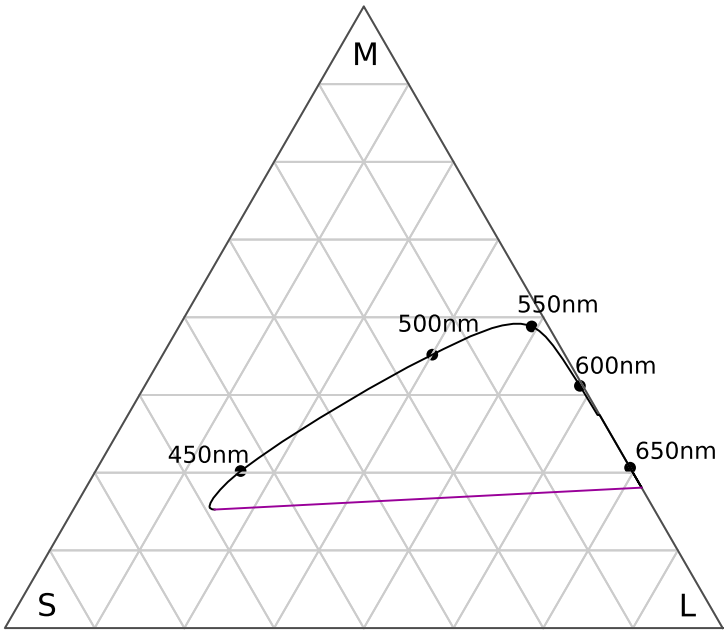

Figure 1.7: Maxwell's triangle. The large triangle represents all possible positive quantum catches, not restricted by what is achievable by actual spectra. The curve with marked wavelengths is the spectral line, this contains all colours made by monochromatic lights - some wavelengths are marked. The line shown in purple is known as the purple line (as it bounds the area containing pinks and purples), this is a boundary too, but no monochromatic light corresponds to it - it is defined by the convex combinations of the two points flanking the concavity in the spectral line between the short and long loci.

been much success in the application of discrimination based measures, but, little success in finding long distance measures. The subject of chapters 3, 4 and 5 is the characterisation, formalisation and application of both discriminative (Weberian) and judgemental (Fechnerian) psychophysical theory by applying information theory to a general model of agency.

1.4 20^{th} (and 21^{st}) Century - Differential Geometry

Beginning with von Helmholtz (1896) and Schrödinger (1920) colour was described using the mathematics of differential geometry, specifically Riemannian geometry - a framework

which at the same time found central importance in the field theories of physics (a famous example of which is Einstein). Using Riemannian geometry, a fairly general (but not completely general!) relationship between a physical parameter and the perception of colour is described. This found a niche when describing perceptual distance: one could get some data about discrimination and then transform a space so that instead of the physical parameter being uniformly spaced, the perceptual space is uniform; euclidean; flat[15]. Doing this may or may not lead to a closed form equation that allows one to express the psychological space in terms of the physical space, but regardless of whether it can be easily written down, such a relationship will always exist.

The Riemannian approach has been the basis for many modern models of colour, most notably the empirical CIE $L^*u^*v^*$colour space (industry standard: referenced as CIE, 1976) and has (fairly) recently been used outside of human colour vision by Vorobyev and Osorio (1998). However, there are still some problems with these spaces. Riemannian geometry is especially good at describing small changes in colour, but when it comes to larger changes such models are found lacking - optimal long range metrics are non-Riemannian (see Backhaus et al. (1984) or even Von der Emde and Ronacher (1994)). Distance between colours are not the standard geodesics of Riemannian geometry, even though geodesics seems to have some relevance (Schrödinger, 1920; Wyszecki and Stiles, 2000).

Yet, Riemannian geometry is only one of many geometries and its use up to now has been more of an empirically tested postulate than theory driven application. It is the purpose of chapter 4 to show where Riemannian geometry is, and is not, a natural form for colour spaces to have.

[15]Flat meaning affine, not meaning without intrinsic curvature. Colour spaces are usually flat in the intrinsic sense

Chapter 2

The Shape of Colour Spaces

"Would it be conceivable for someone to see as black everything that we see as white, and vice versa?"

Ludwig Wittgenstein
Remarks on Colour, McAlister and Schättle trans.

The purpose of this chapter is twofold. First of all, this chapter is intended to give a background to the mathematics that will be used in the later chapters. Colour science has various formalisations for many of the quantities that it uses, being explicit about the formalisation is used is a necessity, especially as the formalisation here is non-standard in some places. Research in colour vision has a human bias, not surprisingly: The implications for cognitive psychology and philosophy of mind are of great concern and *Homo sapien* is the 'model species' in these fields. Here I define quantities that avoid this human bias. The aim here is to provide geometric models without reliance upon any particular photoreceptor absorbances or their number. Some of the work here is part of the day to day business of colour scientists and has been thoroughly studied in humans. The generalisations of these concepts, whilst fairly simple, is either ignored or mistaken.

The second purpose of this chapter is to investigate the similarity between the colour vision of various organisms. This of course relies on the existence of measures which can be equally applied to all organisms. I wish to highlight one similarity in particular – the similarity between 'colourfulness' in various organisms. This concept will be of use later on in this thesis.

The focus here is the large scale geometry of colour spaces. Because the spectral sensitivities of photoreceptors overlap the locus of possible signals is non-trivial - we cannot excite one type of photoreceptor without exciting another, making some values of responses

are impossible to achieve. This chapter explores the geometry of the possible signals and problems related to this geometry.

2.1 Definitions

This section outlines some notation, $\mathcal{C}$ for the convex closure of a set (elsewhere called the hull) and $\mathcal{H}$ for the minimal set of points which when closed is equal to a given convex set (which, in line with usage in computer science, I will call the convex hull). It also explains the inner product notation $\langle \cdot, \cdot \rangle$, which is completely conventional and can be found in many places (e.g. Amari and Nagaoka, 2000). Those familiar with these may wish to skip to section 2.2.

2.1.1 Convexity

The concept of convexity is fundamental to the proceeding chapter. A subset of a vector space is said to be *convex* if for any two points (vectors) in the set that are chosen every point in between them is also in the set. More formally, a convex set $\mathcal{S}$ has the property:

$$\mathbf{x}, \mathbf{y} \in \mathcal{S} \implies (1-k)\mathbf{x} + k\mathbf{y} \in \mathcal{S} \qquad (\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall k \in [0,1]) \tag{2.1.1}$$

The use of convexity in this chapter is of a *set* (in a vector space) and different to the convexity of a *function* used in chapter 4[1]. A convex function is simply a one-to-one function where the set of points above the curve is convex.

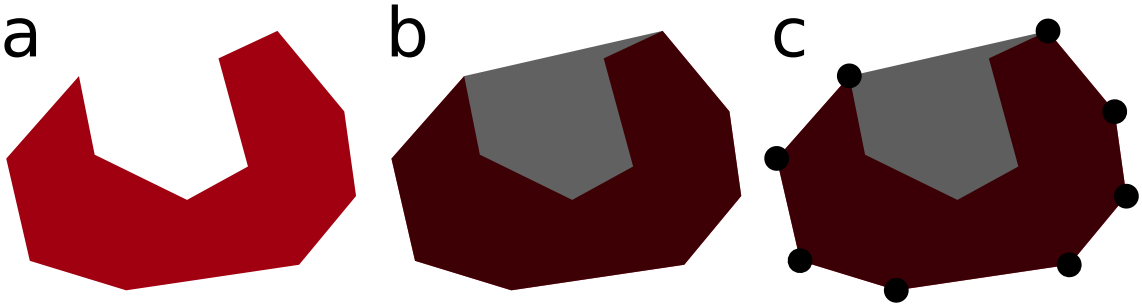

Figure 2.1: Convex constructions: a) A set in $\mathbb{R}^2$ b) The minimal convex covering of the set c) The points in the convex hull.

The definition of convexity naturally leads to the idea of a *(minimal) convex covering*, also known as the *convex closure* of a set of vectors. The convex covering can be defined

[1]Formally, f is a convex *function* iff for some domain X, $\forall x, y \in X, k \in [0,1] : f(kx + (1-k)y) \leq kf(x) + (1-k)f(y)$, it is said to be *strictly* convex if the inequality is strict ($<$ instead of $\leq$)

for all sets of points. This is created by the inclusion of all the convex combinations of points that were not in the original set. The convex covering of $\mathcal{S}$, $\mathcal{C}(\mathcal{S})$, can be written as

$$\mathcal{C}(\mathcal{S}) = \left\{ \sum_{\mathbf{x}\in\mathcal{S}} k_i \mathbf{x} \ : \ \sum_i k_i = 1, k_i \in [0,1] \right\} \tag{2.1.2}$$

In mathematics, this is often called the convex hull, however, in computer science where one must use effecient representations there is another meaning of this term, which is the one I will use here. The *convex hull* of a set is the set of all points in the convex covering of that set which are not convex combinations of other points. My usage here differs from the usual usage, where the convex closure and the hull are one and the same; here I use 'hull' and 'closure' or 'covering' to make the distinction between the set of points needed to describe a convex set and that set itself. The points of the convex hull are the 'extremities' of the set. The convex hull $\mathcal{H}(\mathcal{S})$ of a set $\mathcal{S}$ (in the computer science sense of a minimal representation) is therefore defined as

$$\mathcal{H}(\mathcal{S}) = \{\mathbf{x} \in \mathcal{C}(\mathcal{S}) \ : \ \mathbf{x} \notin \mathcal{C}(\mathcal{S} \setminus \{\mathbf{x}\})\} \tag{2.1.3}$$

it is possible to show, though I shall not do so here, that

$$\mathcal{C}(\mathcal{S}) = \mathcal{C}(\mathcal{H}(\mathcal{S})) \tag{2.1.4}$$

and that:

$$\mathcal{H}(\mathcal{S}) = \mathcal{H}(\mathcal{C}(\mathcal{S})) \tag{2.1.5}$$

$$\mathcal{H}(\mathcal{S}) = \mathcal{H}(\mathcal{H}(\mathcal{S})) \tag{2.1.6}$$

$$\mathcal{C}(\mathcal{S}) = \mathcal{C}(\mathcal{C}(\mathcal{S})) \tag{2.1.7}$$

With equations 2.1.4 though 2.1.7 we can see that any composition of the functions $\mathcal{C}$ and $\mathcal{H}$ is equal to the outermost function.

2.1.2 Inner Products

There are two intimately connected mathematical representations of the spectra of light. One is as a function of wavelength, the other is as a vector containing the intensity at particular wavelengths. If the vector were infinite then the two would be (at least informally) the same. In this spirit, we say that spectra belong to the positive half of an infinite dimensional vector space, informally $[0,\infty)^\infty = \mathbb{R}_+^\infty$.

The two representations are valid because they are both inner product spaces. It is the inner product that really matters. For this reason I shall only use notation for the

inner product, and avoid specific sum/integral representation whenever possible. The inner product between two vectors is given by:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_i \mathbf{A}_i \mathbf{B}_i \tag{2.1.8}$$

There is an integral equivalent. Spectra can be described as functions of wavelength, more specifically, they are densities with respect to a Lebesgue measure $\mathrm{d}\lambda$. All the functions of wavelength discussed here are, in this sense, spectra. For this reason we define the inner product as:

$$\langle a, b \rangle = \langle a(\lambda), b(\lambda) \rangle = \int_\Lambda a(\lambda) b(\lambda) \mathrm{d}\lambda \tag{2.1.9}$$

In this form, the quantum catches $[q_i]$ for a particular incident spectrum are given by

$$q^i = \left\langle w^i, s \right\rangle \tag{2.1.10}$$

can be thought of a linear projection from the infinite dimensional space of spectra to the lower dimensional quantum catch space. This projection is not in general orthogonal as the response functions w^i are not (they overlap to some extent):

$$k^{ij} = \left\langle w^i, w^j \right\rangle \neq 0 \tag{2.1.11}$$

For this reason, quantum catches for an n-chromat inhabit a subspace of $\mathbb{R}^n_+$. This subspace is cone-like with its apex lying on the origin (the 'black point').

2.2 The Spectral Line and Chromaticity Spaces

Before I move on to the main topic of this chapter – the properties of colour solids – I will briefly describe the properties of the more standard notion of chromaticity spaces, and their formulation for a general observer with n photoreceptor classes.

The spectral line is a fundamental object in the study of colour. If we are treating the space of spectra as a vector space it is formed from the projection of basis vectors corresponding to Dirac deltas. Here, I shall write it as the parametric curve $(\gamma(\lambda), \lambda \in \Lambda)$ given by the inner product:

$$\gamma^i(\lambda') = \left\langle w^i(\lambda), \delta(\lambda' - \lambda) \right\rangle = w^i(\lambda') \tag{2.2.1}$$

where δ is the Dirac delta. I include this step to show the trivial relationship between the spectral line and the response functions, which are equal formally, but have different meanings.

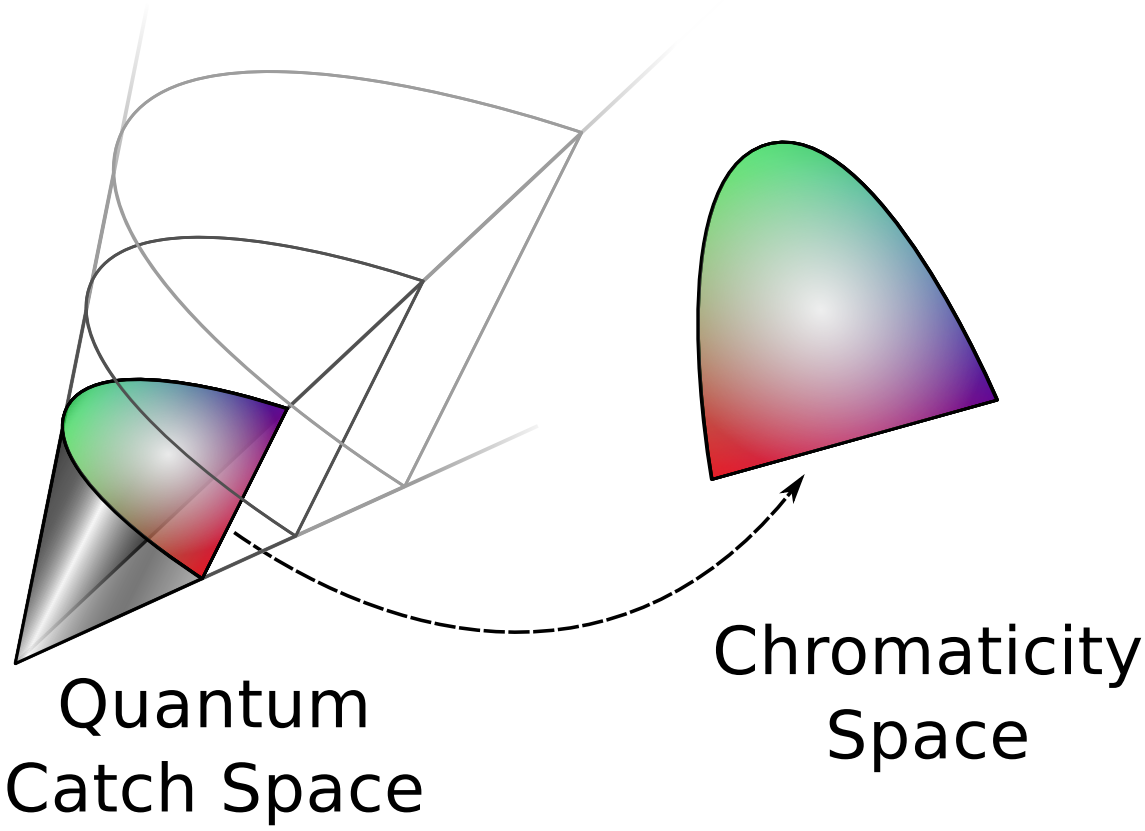

Figure 2.2: The relationship between the quantum catch space and the chromaticity space.

The spectral line can be thought of as the locus of all monochromatic spectra with unit intensity[2]. If we allow arbitrary intensity then the surface:

$$u^i(\lambda, \mu) = \mu\gamma^i(\lambda), \quad \mu \in \mathbb{R}_+ \tag{2.2.2}$$

contains all such spectra. Further, we can demonstrate that convex closure of this surface is equivalent to the set of quantum catches associated with every possible spectrum. The quantum catch of a convex combination of the monochromatic spectra of unit integrated intensity[3] can be written:

$$q^i = \left\langle \gamma^i(\lambda), \bar{s}(\lambda) \right\rangle \tag{2.2.3}$$

with $\bar{s}$ being a member of the collection of normalised spectra, i.e. spectra such that

$$\langle \bar{s}(\lambda), \mathbb{1}(\lambda) \rangle = 1 \tag{2.2.4}$$

Now, if we let spectrum s, of arbitrary intensity, be given by:

$$s(\lambda) = \nu\bar{s}(\lambda) \tag{2.2.5}$$

[2]Though, like the Dirac delta, a completely monochromatic light is only an idealisation

[3]i.e. $\int_\Lambda \bar{s}(\lambda) = 1$.

it will have a quantum catch of:

$$q^i = \left\langle w^i(\lambda), s(\lambda) \right\rangle \tag{2.2.6}$$

which we decompose into 'intensity' ν and a normalised spectrum $\bar{s}$

$$= \left\langle w^i(\lambda), \nu\bar{s}(\lambda) \right\rangle \tag{2.2.7}$$

which is trivially related to the spectral line through equation 2.2.1:

$$= \left\langle \gamma^i(\lambda), \nu\bar{s}(\lambda) \right\rangle = \left\langle \nu\gamma^i(\lambda), \bar{s}(\lambda) \right\rangle \tag{2.2.8}$$

which allows us to define a vector valued function u of wavelength and intensity. u defines the cone in 2.2.

$$= \left\langle u^i(\lambda, \nu), \bar{s}(\lambda) \right\rangle \tag{2.2.9}$$

Now, the operation $\langle \cdot, \bar{s} \rangle$ denotes a convex combination in the sense that each value in $\bar{s}$ can be thought of as a weight in a weighted sum of the values of u and where $\sum \bar{s} = 1$. In other words in set notation we have

$$\{q\} = \left\{ \sum_{\mathbf{x} \in \{u\}} s_i \mathbf{x} \ : \ \sum s_i = 1, s \in [0, 1] \right\} \tag{2.2.10}$$

corresponding to the definition of a convex covering of the set of all values of u (equation 2.1.1). So, we see the quantum catch for every spectrum can be written as a convex combination of points on the surface u (as required).

2.2.1 Chromaticity Space Projection

The cone given by $u(\lambda, \nu)$ is a fundamental object in colour spaces, the projection of it through it's apex gives a class of spaces that describe chromaticity (colour in directions other than its intensity).

Instead of normalising with respect to a uniform spectrum ($\mathbb{1}$). We can uniquely write a spectrum in the form normalised with respect to any spectrum $l(\lambda) > 0$:

$$s(\lambda) = \nu\bar{s}(\lambda), \qquad \langle \bar{s}(\lambda), l(\lambda) \rangle = 1 \tag{2.2.11}$$

. Which implies that

$$\langle s, l \rangle = \nu \tag{2.2.12}$$

We can then consider $l(\lambda)$ to be a generalised measure of luminance. We can also, given the photoreceptor responses, relate $l(\lambda)$ as a luminance vector in quantum catch space

(i.e. as a vector of 'colours', not respresenting a spectrum), $b = [b_i]$, so that b is a solution to

$$\langle w(\lambda), b\rangle = l(\lambda) \tag{2.2.13}$$

Note that b does not uniquely define a given l. A corollary of this is that

$$\nu = \langle s, l\rangle = \langle s, \langle b, w\rangle\rangle = \langle \langle s, w\rangle, b\rangle = \langle q, b\rangle \tag{2.2.14}$$

so:

$$\langle \bar{s}, l\rangle = \left\langle \frac{s}{\nu}, l \right\rangle = \left\langle \frac{s}{\nu}, \langle b, w\rangle \right\rangle = \left\langle \left\langle \frac{s}{\nu}, w^i \right\rangle, b_i \right\rangle = \left\langle \frac{q}{\nu}, b \right\rangle = \left\langle \frac{q}{\langle q, b\rangle}, b \right\rangle \tag{2.2.15}$$

The coordinates $\left[\frac{q^i}{\langle q,b\rangle}\right]$ found on the left hand side of the last inner product are what I will call *embedded chromaticity coordinates*,[4] these have the property that:

$$\left\langle \frac{q}{\langle q, b\rangle}, b \right\rangle = \frac{\langle q, b\rangle}{\langle q, b\rangle} = 1 \tag{2.2.16}$$

Which implies that they describe points within a plane in the quantum catch space. Given that all quantum catches are positive, we can say that not only are they in a plane, but in a simplex of dimension $n-1$. In the case where $b = \mathbb{1}$ this is the unit simplex. The representation of all colours fall within this simplex, in fact they lie within the subset of the simplex

$$\begin{aligned} \frac{q}{\langle q, b\rangle} &\in \mathcal{C}\left(\left\{ \frac{\langle \gamma(\lambda), w\rangle}{\langle \gamma(\lambda), b\rangle} \;:\; \lambda \in \Lambda \right\}\right) \\ &\subseteq \mathcal{C}\left(\left\{ \left[b^i \delta^i_j\right] \;:\; j = 1, 2 \ldots n-1 \right\}\right) \\ &= \text{Simplex}\,(b) \end{aligned} \tag{2.2.17}$$

This states, importantly, that all points in the chromaticity space lie within the convex closure of the chromaticities of the spectral line.

2.2.2 Reducing the Dimension

The embedded chromaticity coordinates are usually projected into a $(n-1)$-dimensional space for efficiency of representation.

I shall call the new coordinates $c = [c^i]$. So we have

$$c = \langle \mathbf{P}_n, q\rangle \tag{2.2.18}$$

[4] I use the term *embedded chromaticity coordinates* as at this point we are still working in an n dimensional space for an n-chromat. We have projected into a plane so have a redundant dimension which is yet to be removed.

where $\mathbf{P}_n$ is a n-by-$(n-1)$ projection matrix. In the commonly used CIE xy space, this is not an isometric projection. The projection is the 2-by-3 matrix:

$$\mathbf{P}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{2.2.19}$$

When comparing different visual systems an isometric transformation is better. Distances and angles are important and isometric projections are well behaved and simply defined for all dimensionalities. This is not a necessity, but 'good practice'. Such transformations are uniquely defined up to rotation. For this, we can choose any matrix that can be expressed as:

$$\mathbf{P} = \mathbf{P}_n\mathbf{K} \tag{2.2.20}$$

where $\mathbf{K}$ is a orthogonal transformation which transforms the luminance vector b to the last basis vector, $\mathbf{e}_n$, so:

$$\mathbb{1}\mathbf{K} = \mathbf{e}_n \tag{2.2.21}$$

and $\mathbf{P}_n$ is a n-by-$(n-1)$ projection matrix:

$$\mathbf{P}_n = \begin{bmatrix} \mathbf{I}_d \\ \mathbb{0} \end{bmatrix} \tag{2.2.22}$$

so that $\mathbf{P}_n\mathbf{e}_n = \mathbb{0}$. In other words, we transform quantum catch coordinates so that the luminance is only in a particular direction and then ignore the corresponding dimension.

2.2.3 The Interpretation of Chromaticity Spaces

Chromaticity spaces are commonplace in colour science and as such are often used as a basis for comparative arguments. It is an unfortunate property of geometric representations of colour that their provision of an intuitive visualisation is also a means by which one may forget that ultimately we require a justification beyond the geometry, and, that the power of a geometric representation is exactly that of the motivation used to produce it.

I mention this because of a widespread misconception in comparative colour vision. This concerns spectral and non spectral colours. We know (very) approximately, that the spectral line traverses the corners of an $(n-1)$-simplex, for example, in human colour vision, it travels approximately from red to blue via green (see figure 1.6), leaving one

edge (red-blue) mostly unvisited. We see similar features in a a tetrachomatic space; in which case there are three unvisited edges, a pentachromatic space where there are six and in an n-chromatic space there are $\frac{1}{2}(n-1)(n-2)$ unvisited lines. The spectral line would follow these edges of a tetrahedron perfectly if the photoreceptor response functions overlapped only with those neighbouring them in the wavelength domain.

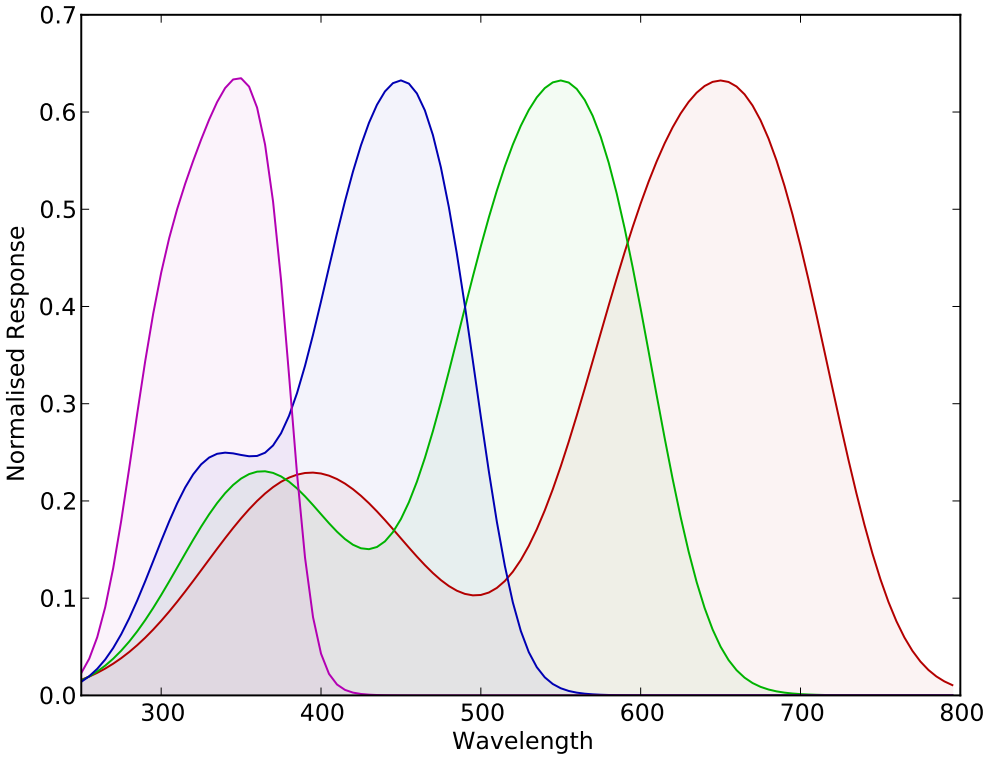

Figure 2.3: Photoreceptor response curves used to make the chromaticity space in figure 2.4.

This quality has lead to an unusual mythology concerning colour spaces, an example of which is found in Thompson (1995, chap 4.) and the references therein. In particular I take issue with numerological statements such as:

> While in a man's chromaticity diagram there is only one intermediate colour [...] namely purple, in tetrachomatic vision there would be three.
>
> *Burkhardt and Maier (1989)*
>
> *See: Thompson (1995)*

A clarification of 'how many' non spectral colours there are on the outside of a colour space is needed.

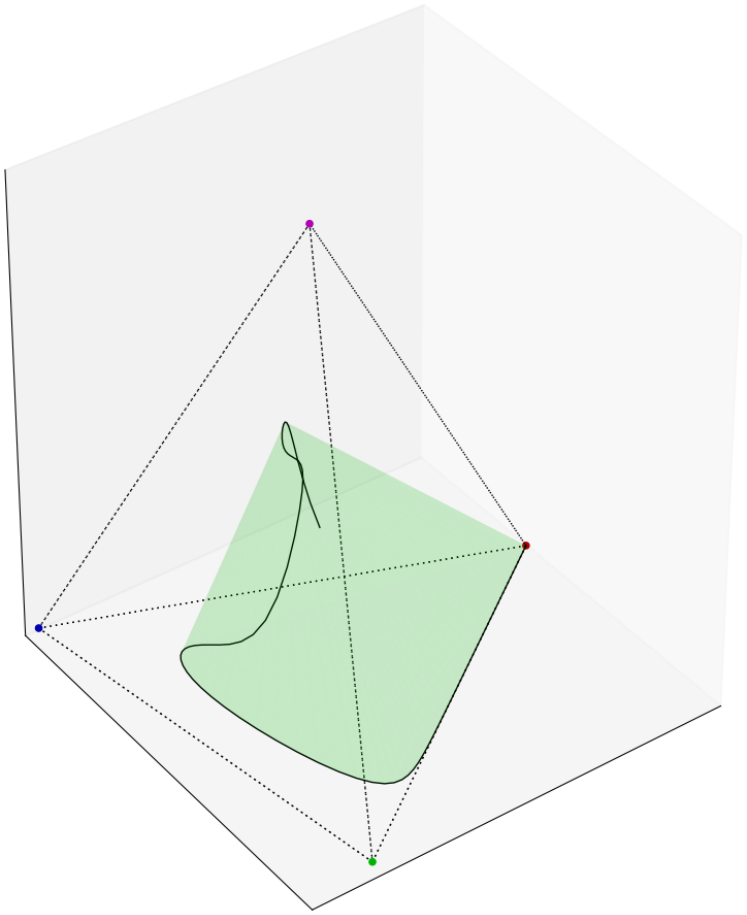

Figure 2.4: The chromaticity space corresponding to the curves in figure 2.3. The green area corresponds to a two dimensional surface of the chromaticity space, just as the curves and lines do in figure 1.6.

The spectral line is always a one dimensional line. The dimensionality of the surface of the n-chromat chromaticity space is always $n-2$ as it is the surface of a bounded, convex volume within a $n-1$ dimensional simplex. It is pretty clear from this that there will be, informally, no non-spectral colours for and dichromats, the same ordinality for trichromats ($n-2=1$), and infinitely more non-spectral colours than spectral lying on the surface when $n \geq 4$. Also, we can see in figure 2.3, that there is no reason to assume that the spectral line will lie on the surface of the chromaticity space in its entirety - the line between blue and purple is non-spectral. The distinction between spectral and non-spectral colours is rather artificial, and is only made so that we can calculate the form of the chromaticity space so that later we can look at where within it spectra lie.

The attraction to thinking about spectral an non-spectral colours is motivated by

human colour vision, where the purple line (the non-spectral locus for trichromats) has special significance - there is a qualitative similarity[5] between it and the spectral line. However, this similarity does not exist for non-trichromatic organisms. I will not use chromaticity spaces in the rest of this thesis, I include them here as they are the most widespread representation of colour and the geometry of their construction reflects much of the geometry in this chapter.

2.3 Colour Solids

The colour solid, or object colour solid, is the collection of quantum catch vectors possible under a constant observer and illumination (Wyszecki and Stiles, 2000).[6]

The quantum catch for a surface with reflectance[7] $r(\lambda)$ under illumination $l(\lambda)$ is given by:

$$q^i = k^i \left\langle w^i, rl \right\rangle = k^i \left\langle r, w^i l \right\rangle = \left\langle r, k^i w^i l \right\rangle \tag{2.3.1}$$

where each k^i is a normalisation constant equal to $1/\left\langle w^i, l \right\rangle$.

The definition of the colour solid is then given by:

$$\mathcal{S}(w) \stackrel{\text{def}}{=} \left\{ \left\langle r, k^i w^i l \right\rangle \ : \ r \in \mathfrak{R} \right\} \tag{2.3.2}$$

where $\mathfrak{R} = [0,1]^\infty$ is the space of all reflectance spectra. This is effectively the linear projection of a hypercube of infinite dimension. It is possible to write $\mathfrak{R}$ as convex combinations of points in $\{0,1\}^\infty = \mathfrak{T}$. In other words, $\mathfrak{T}$ is the set of all spectra with either complete or zero reflectance at any given wavelength, from which all reflectances can be produced by convex combination (a weighted and normalised sum). $\mathfrak{T}$ can be interpreted as the set of all vertices of the infinite dimensional hypercube containing all reflectances, $\mathfrak{R}$. The geometric interpretation of $\mathfrak{R}$ and $\mathfrak{T}$ should make it clear that $\mathfrak{R}$ is the convex covering of $\mathfrak{T}$:

$$\mathfrak{R} = \mathcal{C}(\mathfrak{T}) \tag{2.3.3}$$

and that $\mathfrak{T}$ is the convex hull of the reflectances:

$$\mathfrak{T} = \mathcal{H}(\mathfrak{R}) \tag{2.3.4}$$

[5] We can use the same Lebesgue measure.

[6] Here, for simplicity, the colour solid of an n-chromat is normalised to lie within the unit n-cube.

[7] Reflectance spectra are bounded above and below, so $r \in [0,1]^\infty$.

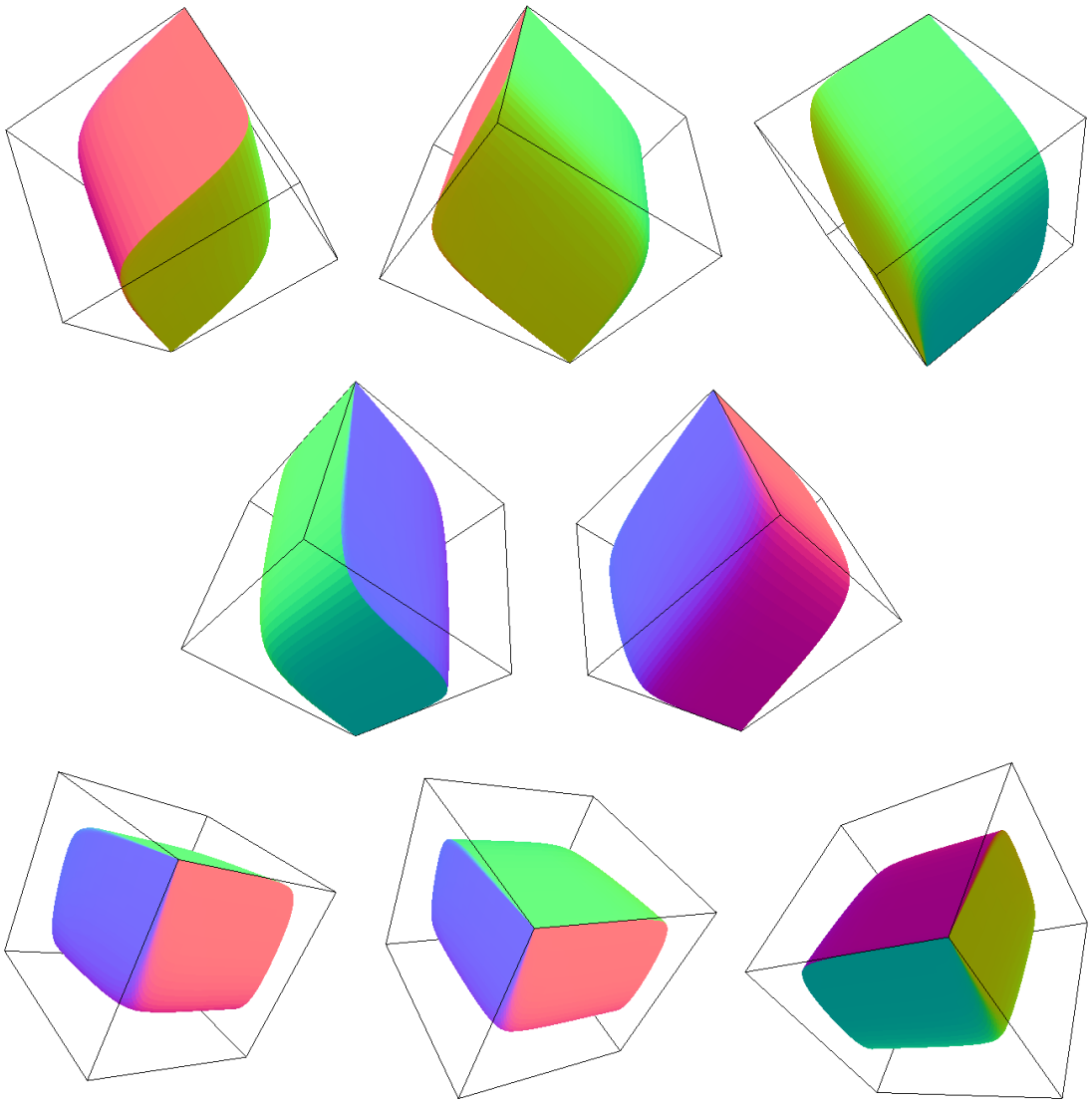

Figure 2.5: Various views of the honeybee colour solid. It is coloured according to the direction of vectors normal to the surface, not by actual colour, although there is some qualitative correspondence. Here, like in the rest of this section, I assume that the illumination spectrum is uniform.

Before I continue, I need to mention the action of projections upon convexity. As mentioned above, a point $\mathbf{b}$ is convex combination of $\mathbf{a}$ and $\mathbf{c}$ iff $\mathbf{b} = k\mathbf{a} + (1-k)\mathbf{c}, k \in [0,1]$. Now, if we take the linear projection under $\mathbf{P}$ of $\mathbf{b}$ we see that:

$$\mathbf{Pb} = \mathbf{P}\left(k\mathbf{a} + (1-k)\mathbf{c}\right) = k(\mathbf{Pa}) + (1-k)(\mathbf{Pc}) \tag{2.3.5}$$

in other words, the projections of convex combinations of points are the convex combination of their projections. This means that we have the following relations for the convex

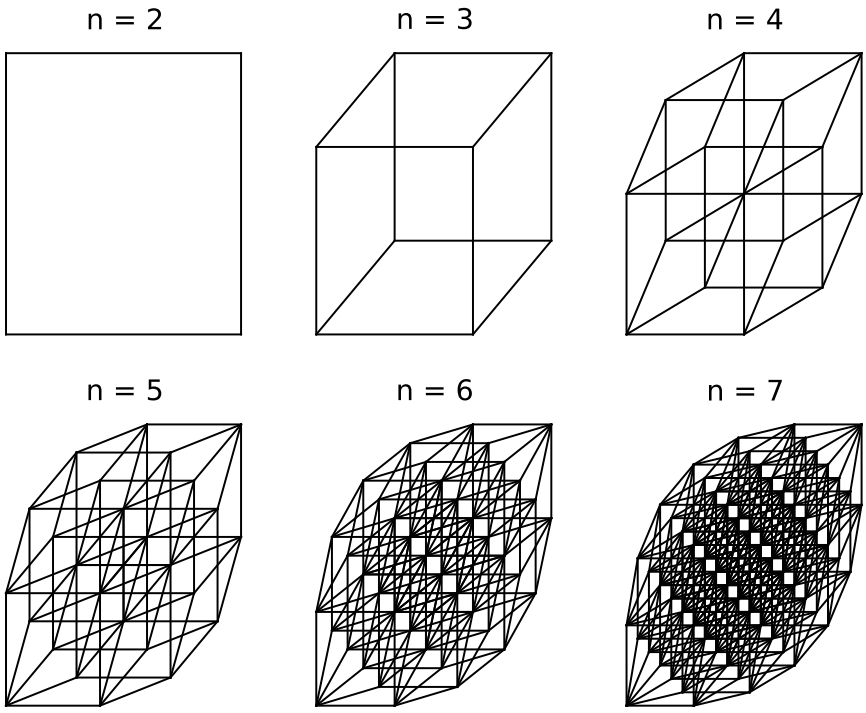

Figure 2.6: Projections of hypercubes of increasing dimension. With the appropriate projection, the outer boundary forms a colour solid. n is the dimensionality of the hypercube. As we increase the dimension of the cube our projection becomes increasingly similar to a colour solid in appearance.

covering and hull:

$$\{\mathbf{Py} \ : \ \mathbf{y} \in \mathcal{C}(\mathcal{X})\} = \mathcal{C}(\{\mathbf{Px} \ : \ \mathbf{x} \in \mathcal{X}\}) \tag{2.3.6}$$

$$\{\mathbf{Py} \ : \ \mathbf{y} \in \mathcal{H}(\mathcal{X})\} = \mathcal{H}(\{\mathbf{Px} \ : \ \mathbf{x} \in \mathcal{X}\}) \tag{2.3.7}$$

From the definition in equation 2.3.2 we see that this implies that the colour solid for a given set of photoreceptor responses w and illumination $\mathcal{S}(w, l)$:

$$\begin{aligned} \mathcal{S}(w,l) &= \left\{\left\langle r, k^i w^i l\right\rangle \ : \ r \in \mathcal{C}(\mathfrak{T})\right\} \\ &= \mathcal{C}\left(\left\{\left\langle t, k^i w^i l\right\rangle \ : \ t \in \mathfrak{T}\right\}\right) \\ &= \mathcal{C}\left(\mathcal{H}\left(\left\{\left\langle t, k^i w^i l\right\rangle \ : \ t \in \mathfrak{T}\right\}\right)\right) \end{aligned} \tag{2.3.8}$$

so also

$$\mathcal{H}(\mathcal{S}(w)) = \mathcal{H}\left(\left\{\left\langle t, k^i w^i l\right\rangle \ : \ t \in \mathfrak{T}\right\}\right) \tag{2.3.9}$$

The latter is better to work with as $\mathcal{H}(\mathcal{S}(w))$ is approximated numerically as a finite collection of points, whereas the corresponding solid $(\mathcal{S}(w))$ is, except from some pathological cases[8], infinite – even when its hull is finite. We work with the hull of the colour solid, as it is a unique representation of the solid which is numerically tractable.

2.3.1 Extreme Spectra

Often the colour solid is calculated using a set of so called 'extreme spectra' (Logvinenko, 2009; Vorobyev, 2003). I shall call the set of these spectra $\mathfrak{X}(n)$. A particular extreme spectrum is parametrised by $n-1$ real values, which I will call λ^i and a binary value $\kappa \in \{-1, 1\}$.

$$\eta(\lambda; \lambda^1 \ldots \lambda^{n-1}, \kappa) \stackrel{\text{def}}{=} \tfrac{1}{2} + \tfrac{\kappa}{2} \prod_{i=1}^{n-1} \Phi\left(\lambda - \lambda^i\right) \tag{2.3.10}$$

where $\Phi(x) = 1 - 2\Theta(x)$ and $\Theta(\cdot)$ is the step function defined by:

$$\Theta(x) \stackrel{\text{def}}{=} \begin{cases} 0, & x < 0 \\ 1, & \textit{otherwise} \end{cases} \tag{2.3.11}$$

The entire set of these is:

$$\mathfrak{X}(n) \stackrel{\text{def}}{=} \left\{\eta(\lambda; \lambda^1 \ldots \lambda^{n-1}, \kappa) \;:\; \lambda^i \in \Lambda,\ \lambda^i > \lambda^{i+1},\ \kappa \in \{-1, 1\}\right\} \tag{2.3.12}$$

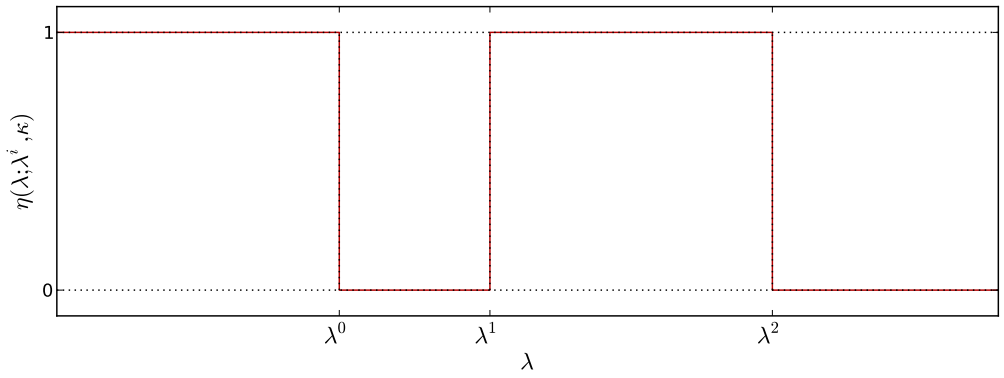

Figure 2.7: An example of an extreme spectrum for a tetrachromat. In this case there are three step changes, at λ^1, λ^2 and λ^3. Changing κ has the result of reflecting the spectrum through the line where $\eta = 0.5$.

The colour solid for humans can be calculated from the extreme spectra. The two values of κ define two $(n-1)$-dimensional surfaces. Such calculations are correct when the

[8] Only when the hull is a single point is the convex covering finite

spectral line in the chromaticity space belongs entirely to its convex hull (Luther, 1927; Nyberg, 1928). When this is not the case, other techniques must be used.

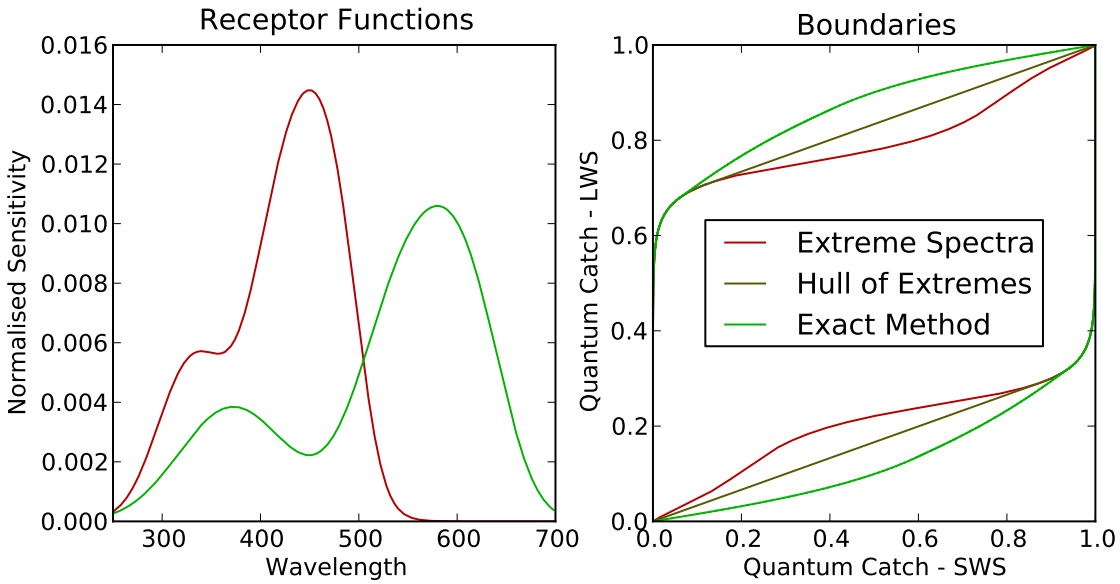

Figure 2.8: Examples of colour solids calculated with three different methods. The extreme spectrum method produces a concave shape. Taking the convex hull remedies this somewhat, but still does not calculate the exact result.

We can see the deficiency of the extreme spectra method in figure 2.8. In this simple case, the actual colour solid extends far beyond the boundary as calculated using the extreme spectra. This can be problematic in two ways, firstly the resulting solid formed by extreme spectra calculations may not be convex, leading to problems for any result that requires convexity. Secondly, the extreme spectrum method works best when the sensitivity functions are uni-modal, so if this method is used to compare uni-modal sensitivities with non-uni-modal sensitivities a distorted comparison is made. This is a problem in the comparative work, such as Vorobyev (2003), where the presence of pigmented oil droplets causes the spectra to narrow, causing an increase in the volume of the colour solid (used as a measure of the 'number of colours that can be distinguished'), but also an accompanying decrease in the underestimate of made by the extreme spectrum method. It is a systematic error which overestimates the value of filtering. Thus, we can be certain that the numerical results of Vorobyev (2003) are an overestimate of the benefit (in terms of discrimination) of coloured oil droplets.

The numerical calculation of the full colour solid without resorting to extreme spectra or linear programming can be found in appendix E.3.

2.3.2 Symmetry

The colour solid has a point of reflectional symmetry at $(\frac{1}{2}, \frac{1}{2} \cdots)$. This is due to the symmetry of the set of reflectances it projects from. We can see this by first noting that the set of reflectance spectra $\mathfrak{R}$ has the following property:[9]

$$\forall s \in \mathbb{R}^\infty : \tfrac{1}{2} + s \in \mathfrak{R} \iff \tfrac{1}{2} - s \in \mathfrak{R} \tag{2.3.13}$$

i.e. if we conceptualise a hypercube representing $\mathfrak{R}$, it is centrally symmetric around $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2} \cdots)$.

The symmetry of the colour solid motivates a parametrisation of it in which the symmetry is easily represented.

2.3.3 Metamers Sets and Their Volumes

Because of the projective nature of colour spaces, it is possible for spectra to be different and be projected onto the same point of a given colour solid. This phenomenon is called metamerism and has been well studied (see e.g. Wyszecki and Stiles, 2000). Usually the set of spectra under consideration are reflectance spectra and the colour space is the colour solid. I wish now to briefly mention a few properties of the colour solid and its corresponding metameric spectra. I shall consider the properties of $[0,1]^n$, which is equal to $\mathfrak{R}$ in the limit of $n \to \infty$.

I would like to introduce some notation. First of all I shall call the (not necessarily infinite) space of spectra sampled at m points: $\mathcal{R} = [0,1]^m$. The number of cone classes possessed by an organism is denoted n. The set of all spectra which can produce a colour q – i.e. the metamer set of $\mathcal{R}$ at q – is defined as:

$$\mathcal{M}(w, q) = \{s \in \mathcal{R} \ : \ \langle s, w \rangle = q\} \tag{2.3.14}$$

This can be related to an intuitive notion of an $(m-n)$-dimensional volume, $V(\mathcal{M})$. I take this volume to be given by the Lebesgue measure on $\mathbb{R}^{m-n}$. Informally, if there are more spectra in $\mathcal{M}_1$ than $\mathcal{M}_2$, then $V(\mathcal{M}_1) > V(\mathcal{M}_2)$.

Let's now concentrate on a line, γ, that passes through the centre of symmetry of the colour solid in the direction of a unit vector v, parametrised by k:

$$\gamma^i(k) = kv^i + \tfrac{1}{2} \tag{2.3.15}$$

[9] For each wavelength, if the reflectance takes a value of $x \in [0,1]$ then there is another in $\mathfrak{R}$ that for that wavelength has the value $1-x$. i.e. $x \in [0,1] \implies 1-x \in [0,1]$

From the symmetry of the colour solid we have:

$$\mathcal{M}\left(w,\gamma(k)\right)=\left\{\tfrac{1}{2}-m\ :\ \ m\in\mathcal{M}\left(w,\gamma(-k)\right)\right\} \tag{2.3.16}$$

a corollary of which is

$$V(\mathcal{M}\left(w,\gamma(k)\right))=V(\mathcal{M}\left(w,\gamma(-k)\right)) \tag{2.3.17}$$

At the centre of symmetry the volume of the metamer set is nonzero:

$$V\left(\mathcal{M}\left(w,\gamma(0)\right)\right)>0 \tag{2.3.18}$$

and at some value of k we go beyond the boundary of the solid and $\mathcal{M}=\varnothing$. This means that the volume of metamer set *is* zero:

$$\exists k : V(\mathcal{M}\left(w,\gamma(k)\right))=0 \tag{2.3.19}$$

Now, considering three points along the line, where k is equal to $-k_1$, k_2 or k_1. so that $-k_1\leq k_2\leq k_1$, we define the convex combination

$$\mathcal{N}=\left\{\frac{(k_2-k_1)m_0+(k_1-k_2)m_2}{2k_1}\ :\ \begin{array}{l}\forall m_0\in\mathcal{M}\left(w,\gamma(-k_1)\right)\\ \forall m_2\in\mathcal{M}\left(w,\gamma(k_1)\right)\end{array}\right\} \tag{2.3.20}$$

which is the same or greater in volume as the two sets it is derived from. We can see this by considering only a single metamer from one of the sets, say $m\in\mathcal{M}\left(w,\gamma(k_1)\right)$, so:

$$\mathcal{N}'=\left\{\frac{(k_2-k_1)m_0+(k_1-k_2)m}{2k_1}\ :\ \forall m_0\in\mathcal{M}\left(w,\gamma(k_1)\right)\right\} \tag{2.3.21}$$

and $V(\mathcal{N}')=V(\mathcal{M})$ as this kind of linear combination of spectra cannot transform unique spectra to identical ones. However, some convex combinations of spectra in $\mathcal{M}\left(w,\gamma(k_1)\right)$ and $\mathcal{M}\left(w,\gamma(-k_1)\right)$ may be the same, but there will always be at least the spectra in $\mathcal{N}'$. $\mathcal{N}$ must be an improper subset of the 'actual' metamer set at k_2:

$$\mathcal{N}\subseteq\mathcal{M}\left(w,\gamma(k_2)\right) \tag{2.3.22}$$

and therefore:

$$V(\mathcal{N})\leq V(\mathcal{M}\left(w,\gamma(k_2)\right)) \tag{2.3.23}$$

We can restate this in words: given a point in the solid, there is a set $\mathcal{N}$ made constructed from the metamer sets at opposite sides of the solid, which is definitely not bigger than the metamer set at that point. The set is definitely not bigger than those it is constructed

from either. The metamer volume is then bigger than that of $\mathcal{N}$ and thus bigger than the two metamer sets from which $\mathcal{N}$ was constructed. So,

$$V(w, \gamma(k_2)) \geq V(w, \gamma(k_1)) \tag{2.3.24}$$

Remembering the symmetry of the colour solid around the centre, $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2} \ldots)$, we can infer that the centre is the point where the volume is largest. This is true no matter what the functions w^i happen to be.

This result can be easily generalised to any linear projection of any set of points that is closed under convexity, but I will not go down that route here.

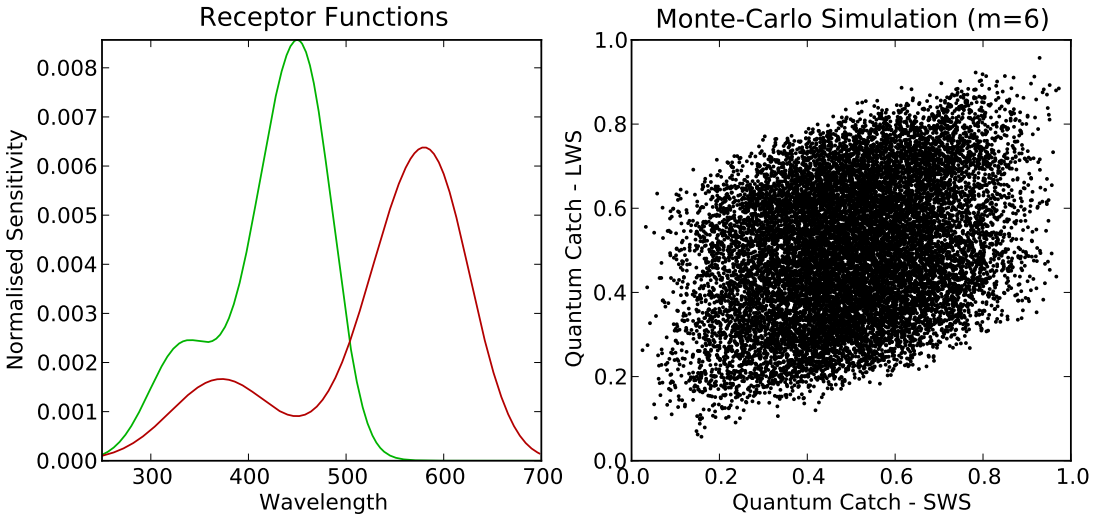

Figure 2.9: Monte-Carlo Sampling of $\mathcal{R}$ for a dichromat with A1 type pigments (Stavenga et al., 1993) under spectrally uniform illumination. $\lambda_{\max}$ values for SWS and LWS sensitive cones are $450nm$ and $580nm$ respectively.

2.3.4 The Interpretation of Metamer Set Volumes

Clearly, not every spectrum in $[0,1]^\infty$, or indeed $[0,1]^m$, exists in the real world. This would seem to be a problem for the application of the theory above. I will go into the details of this in chapter 6. For now though, let us consider what it is that was shown in the previous section. The proof works by taking a *process* - convex combination - and showing how the geometric structure is induced by it. Although it is possible to think of the spectra existing as actual spectra, a more useful conception is of them as *possibilities*. Let me unpack this a little. If I am presented with a colour from the middle of the colour solid, there are many *possible* ways of using the *process* of mixing (convex combination) to achieve that colour. If I choose a colour exactly at the edge of the colour solid, then

there is no way I can mix other colours to achieve it. In other words, I *do not claim* that the volumes of metamer sets of $[0,1]^m$ should be interpreted as statistical features of natural scenes (though there is some correspondence which can be seen in Nascimento et al. 2002), but as a model that corresponds, at least qualitatively, to the number of physical explanations for a given colour. As I have indicated, I will present a fuller argument in chapter 6.

2.4 Comparative Colour Solids

Using the geometry of colour solids it is possible to analyse the relationship between the perceptual spaces of different organisms. This allows us to come up with general rules about comparative colour vision and motivates a class of coordinate systems that are particularly relevant to such studies. With the argument here, for any point in the colour solid of one organism we can find a convex volume in the other organisms colour solid that *must* contain the colour of the underlying spectrum. The concepts I use here are formally the same as those used in the study of metamerism. However, the study of metamerism is usually concerned with the occurrence of spectra in nature that don't produce the same colours to humans under different illuminations (Alsam and Finlayson, 2007; Feng and Foster, 2012; Finlayson and Morovic, 2004; Ohta and Wyszecki, 1975; Trussell, 1991), whereas here I am concerned with the *a priori* relationship between observers with different photoreceptor classes.

The question I ask here is "what does the colour for one organism tell me about the colour for another".

2.4.1 The Monochromatic Case

I will begin with the comparison of two monochromats. As we find with many multidimensional problems it is easier to begin with the one dimensional case and work up. For this problem, it is the only case that it is simple to visualise.

We begin by noting that for two monochromats we can construct a two dimensional colour solid as if the observer's individual photoreceptor types belonged to the same observer. This is perfectly a valid thing to do as a colour solid is, essentially, a diagram that shows where there is a spectrum that fulfils the constraints of yielding a given quantum catch in each photoreceptor class. Similarly, we can ask if

Let us call the quantum catch for one observer a and the other b. Then for a colour a, we can find an interval in $[b_1, b_2]$ where there is spectrum that corresponds to both a and

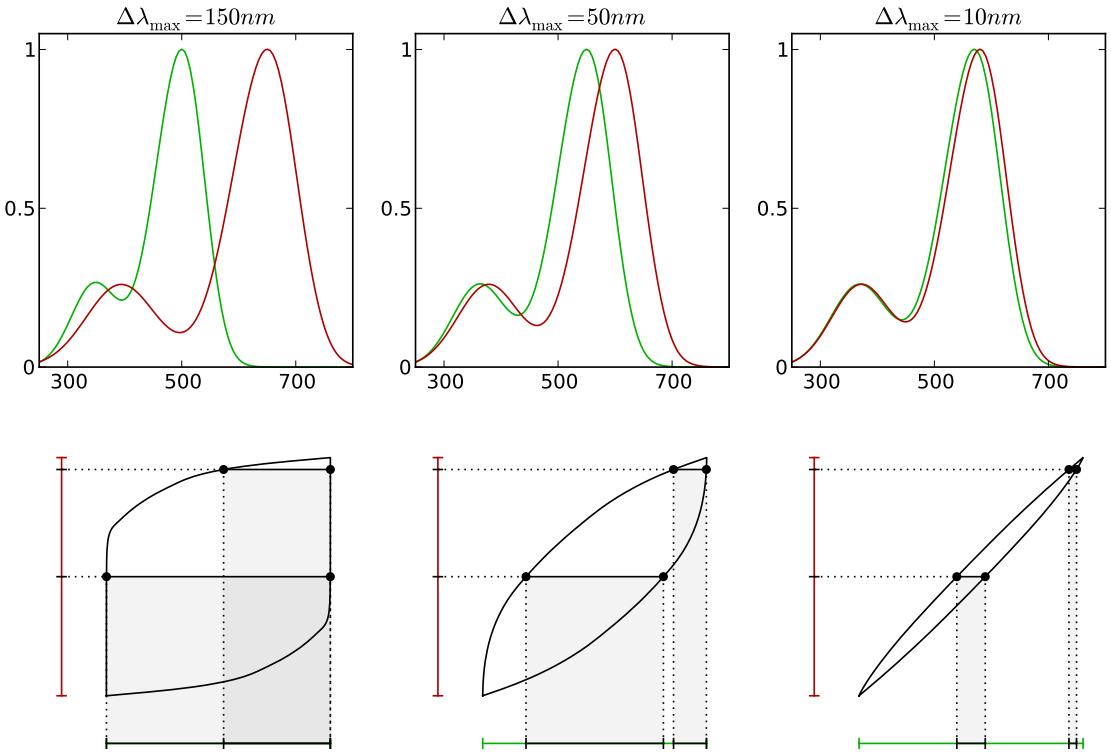

Figure 2.10: The ability to predict the colour for one (green) organism from that of another (red) is related to the colourfulness and the similarity of the photoreceptor functions. Similarity between photoreceptors (the inner product between the two would be a suitable measure for this) narrows the colour solid, which is always broadest at $\frac{1}{2}$.

$b \in [b_1, b_2]$, this is shown in in figure 2.10. The size of the interval is a measure of how well we can predict b from a. As the colour solid is convex, with reflection symmetry about the point $(\frac{1}{2}, \frac{1}{2})$ and contains the points $(0, 0)$ and $(1, 1)$ we can show that it is broadest at $a = \frac{1}{2}$. If, for the two dimensional solid we take $a = f(b)$ to be the upper bounding curve (where $a \geq b$) then the width, $w(a)$, of the interval $[b_1, b_2]$ at a is given by:

$$\begin{aligned} b_1 &= f(a) \\ b_2 &= 1 - f(1-a) \\ w(a) = b_2 - b_1 &= f(a) + f(1-a) - 1 \end{aligned} \tag{2.4.1}$$

We can express the convexity of the solid as

$$kf(x) + (1-k)f(y) \leq f(kx + (1-k)y) \tag{2.4.2}$$

Taking $k = \frac{1}{2}$ we have

$$\tfrac{1}{2}(f(x) + f(y)) \leq f\left(\tfrac{1}{2}(x+y)\right) \tag{2.4.3}$$

and looking at the points where $x = a$ and $y = 1 - a$ we get:

$$\begin{aligned} f(a) + f(1-a) &\leq 2f\left(\tfrac{1}{2}\right) \\ w(a) + 1 &\leq w\left(\tfrac{1}{2}\right) + 1 \\ w(a) &\leq w\left(\tfrac{1}{2}\right) \end{aligned} \tag{2.4.4}$$

demonstrating the interval is indeed widest at $a = \frac{1}{2}$.

More generally, we can take a new parametrisation where $a = \beta + \frac{1}{2}$ for $a \geq \frac{1}{2}$ and show that for $\Delta\beta \geq 0$ that:

$$w\left(\tfrac{1}{2} + \beta + \Delta\beta\right) \leq w\left(\tfrac{1}{2} + \beta\right) \tag{2.4.5}$$

Firstly we note that $\frac{1}{2}+\beta$ and $\frac{1}{2}-\beta$ are convex combinations of $\frac{1}{2}+\beta+\Delta\beta$ and $\frac{1}{2}-\beta-\Delta\beta$. So we can write:

$$\begin{aligned} \tfrac{1}{2} - \beta &= k_1\left(\tfrac{1}{2} + \beta + \Delta\beta\right) + (1 - k_1)\left(\tfrac{1}{2} - \beta - \Delta\beta\right) \\ k_1 &= \frac{\beta + \frac{1}{2}\Delta\beta}{\beta + \Delta\beta} \end{aligned} \tag{2.4.6}$$

$$\begin{aligned} \tfrac{1}{2} + \beta &= k_2\left(\tfrac{1}{2} + \beta + \Delta\beta\right) + (1 - k_2)\left(\tfrac{1}{2} - \beta - \Delta\beta\right) \\ k_2 &= \frac{\frac{1}{2}\Delta\beta}{\beta + \Delta\beta} \end{aligned} \tag{2.4.7}$$

this relies on the positivity of $\Delta\beta$. Then from equation 2.4.2 we can write:

$$\begin{aligned} w\left(\tfrac{1}{2} + \beta\right) + 1 &= f\left(\tfrac{1}{2} + \beta\right) + f\left(\tfrac{1}{2} - \beta\right) \\ &\geq k_1 f\left(\tfrac{1}{2} + \beta + \Delta\beta\right) + (1 - k_1) f\left(\tfrac{1}{2} - \beta - \Delta\beta\right) \\ &\quad + k_2 f\left(\tfrac{1}{2} + \beta + \Delta\beta\right) + (1 - k_2) f\left(\tfrac{1}{2} - \beta - \Delta\beta\right) \end{aligned} \tag{2.4.8}$$

It is easy to see that $k_1 + k_2 = 1$ so:

$$\begin{aligned} w\left(\tfrac{1}{2} + \beta\right) + 1 &\geq f\left(\tfrac{1}{2} + \beta + \Delta\beta\right) + f\left(\tfrac{1}{2} - \beta - \Delta\beta\right) \\ &= w\left(\tfrac{1}{2} + \beta + \Delta\beta\right) + 1 \end{aligned} \tag{2.4.9}$$

and

$$w\left(\tfrac{1}{2} + \beta\right) \geq w\left(\tfrac{1}{2} + \beta + \Delta\beta\right) \tag{2.4.10}$$

as required. So, for one monochromat, the ability to predict the colour for a different monochromat increases with distance from $\frac{1}{2}$.

2.4.2 Polychromatic Cases

We have a very similar situation for dichromats, trichromats and n-chromats in general. We simply make a $2n$-dimensional colour solid, and cut it with an n dimensional hyperplane the corresponds to a fixed colour for one observer, leaving a n-volume in the solid of the other. Indeed, there is no reason why we cannot use observers with different numbers of photoreceptor classes.

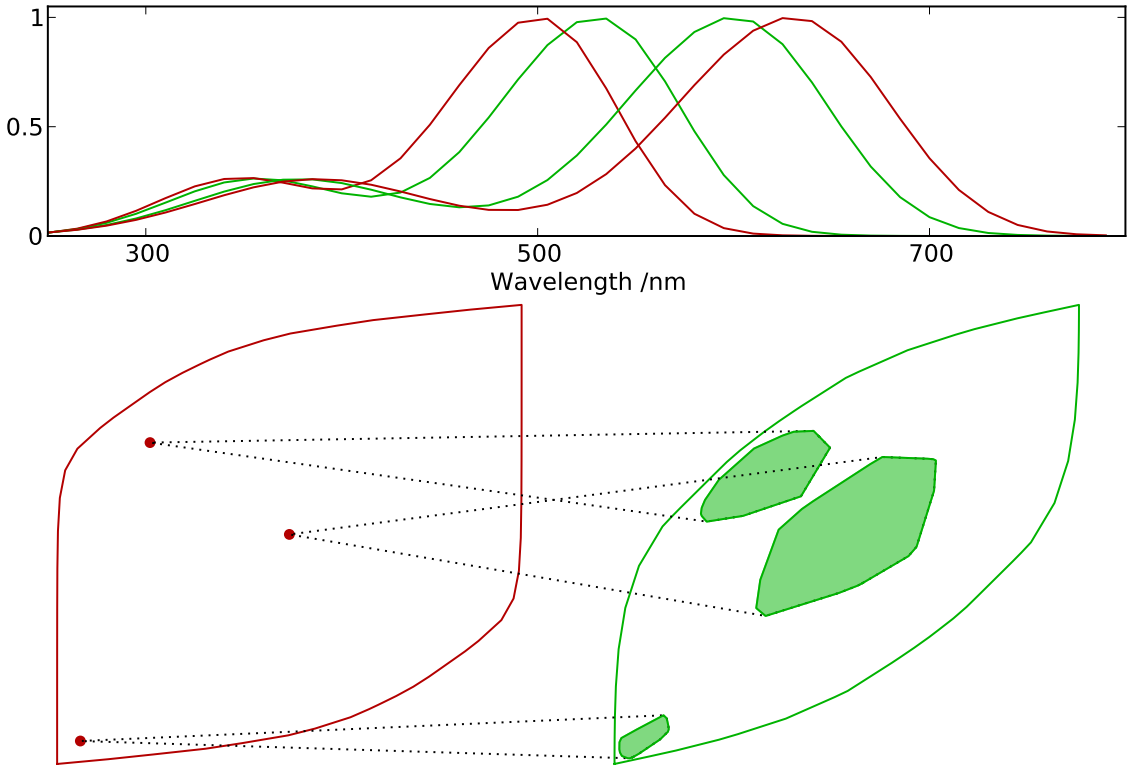

Figure 2.11: The relationship between the colour solids of two dichromats. The curves at the top correspond to the response curve of each photoreceptor class, with the set of red curves corresponding to one organism and green to another. The colour solids at the bottom are coloured correspondingly. We see that in this case each point in one space corresponds to an area in the other. I have chosen to display three points, one which is almost black, one which is grey and one which is very colourful. Like in figure 2.10, it is clear that the distance from the centre of a point in one solid is an indication of the size of the area to which it corresponds in the other.

We can see in figure 2.11 that again, that colours lying in the centre of the colour solid of one observer are the worst at predicting the colour for another.[10] Generally, the further

[10]I do not provide a proof of this like in the monochromatic case, the proof should be basically the same but more lengthy.

one gets to the edge of the solid, the more sure we can be about the location of the colour in another observers colour solid. If we had not demonstrated that the extreme colours, $\mathfrak{X}(n)$, did not necessarily lie on the surface of the colour solid, then one might be led to believe that there was a bijective mapping from all the points on the surface of one colour solid, to those on another. This is not true, the only determined mapping points that we can be completely sure of without knowing the details of the photoreceptor responses are those of black and white.

2.4.3 Summary of Comparing Colour Solids

Essentially, the size of the metamer set places a restriction on the predictive power from one observer's colour to another's. When the metamer set is big, it is difficult to make predictions, when it is small, it is easy. As there are 'more' metamers near the center of the colour solid, the colours near the middle of the solid provide less predictive power. As I have said before, there are no direct biological consequences of this. However, if we view this in terms of the number of processes that can provide mixtures (instead of *spectra* that can be combined) we do obtain something biologically relevant. Indeed, I have phrased this section in terms of prediction in an idealised system, this only serves to highlight the stucture of the solid and it's relationship with convexity – namely, that convex combinations provide a means to establish a natural notion of colourfulness, as a distance from the centre of the colour solid.

2.5 Colourfulness

The discussion so far motivates a coordinate system for the object colour system. In the human case it is equivalent to that of Logvinenko (2009), although his motivations are very different.[11]

2.5.1 Definition

There are four quantification of colourfulness that I would like to define here. They are all simple geometric distances, all of them very similar to each other. They differ in whether they a distances measured from the centre of symmetry or from the line between the black and white points, and, whether or not they are corrected so that they are unity at the boundary of the colour solid.

[11] His aim was to use extreme spectra for a spectral description of colours, something which is impossible in the general (i.e. not human) case, as I have discussed.

They have the general form (I use bold type to denote vectors):

$$J(\mathbf{q}) \stackrel{\text{def}}{=} \frac{\| \mathbf{q} - g(\mathbf{q}) \|}{f(\mathbf{q})} \tag{2.5.1}$$

where f and g are specific to the particular quantity (see table 2.1).

Colourfulness	**Symbol**	$f(\mathbf{q})$	$g(\mathbf{q})$
Spherical	J_R	1	$\frac{1}{2}$
Cylindrical	J_C	1	$\mathbb{1} \langle \mathbb{1}, \mathbf{q} \rangle$
Corrected Spherical	$\bar{J}_R$	$h(\mathbf{q} - \frac{1}{2}, \frac{1}{2})$	$\frac{1}{2}$
Corrected Cylindrical	$\bar{J}_C$	$h(\mathbb{1} \langle \mathbb{1}, \mathbf{q} \rangle - \mathbf{q}, \mathbb{1} \langle \mathbb{1}, \mathbf{q} \rangle)$	$\mathbb{1} \langle \mathbb{1}, \mathbf{q} \rangle$

Table 2.1: The functions f and g for different quantifications of colourfulness. Here I assume that the one vector is normalised, i.e. $\| \mathbb{1} \| = 1$.

I define the function h so that for an n chromat:

$$h(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \sup \left\{ k : \mathbf{a} + \frac{k(\mathbf{b} - \mathbf{a})}{\| \mathbf{b} - \mathbf{a} \|} \in \mathcal{S}(w) \right\} \tag{2.5.2}$$

In words, $h(a, b)$ finds the distance to the furthest point on the colour solid in direction of a from point b.

The two uncorrected quantities are simply the euclidean distance from the centre of the colour solid, in the case of the spherical colourfulness J_R, and the euclidean distance in the null space of the one-vector in the case of cylindrical colourfulness J_C.

Details of Figure 2.12

The colour solid is formed by the $n-1$ dimensional surfaces (in 2D – curves) labelled Ψ and Ω. It lies within the unit n-cube (a square) $\square OSWL$. O, the black point and origin, lies at one tip of the solid and W, the white point, at the other. G, the grey point and centre of symmetry lies half way between O and W and is the centre of the $(n-1)$-sphere (circle) circumscribing O, S, W, L and I. This sphere is the locus of all points with spherical colourfulness equal to the maximum (whether physically achievable or not). An $(n-1)$, 1-cylinder (square) shown as $\square wxyz$ inscribes the sphere (circle) so that the black and white points are at the the centre of the capping $(n-2)$-spheres (line segments $\overline{wx}$ and $\overline{yz}$). The sides of the cylinder lie on the locus of points with cylindrical colourfulness equal to the maximum for any visual system. These sides (in 2D, line segments $\overline{xy}$ and $\overline{zw}$) are tangent to the sphere halfway between the cylinders caps (here, at L and S). The spherical colourfulness of the point Q is given by the length of the line $\overline{GQ}$ – The

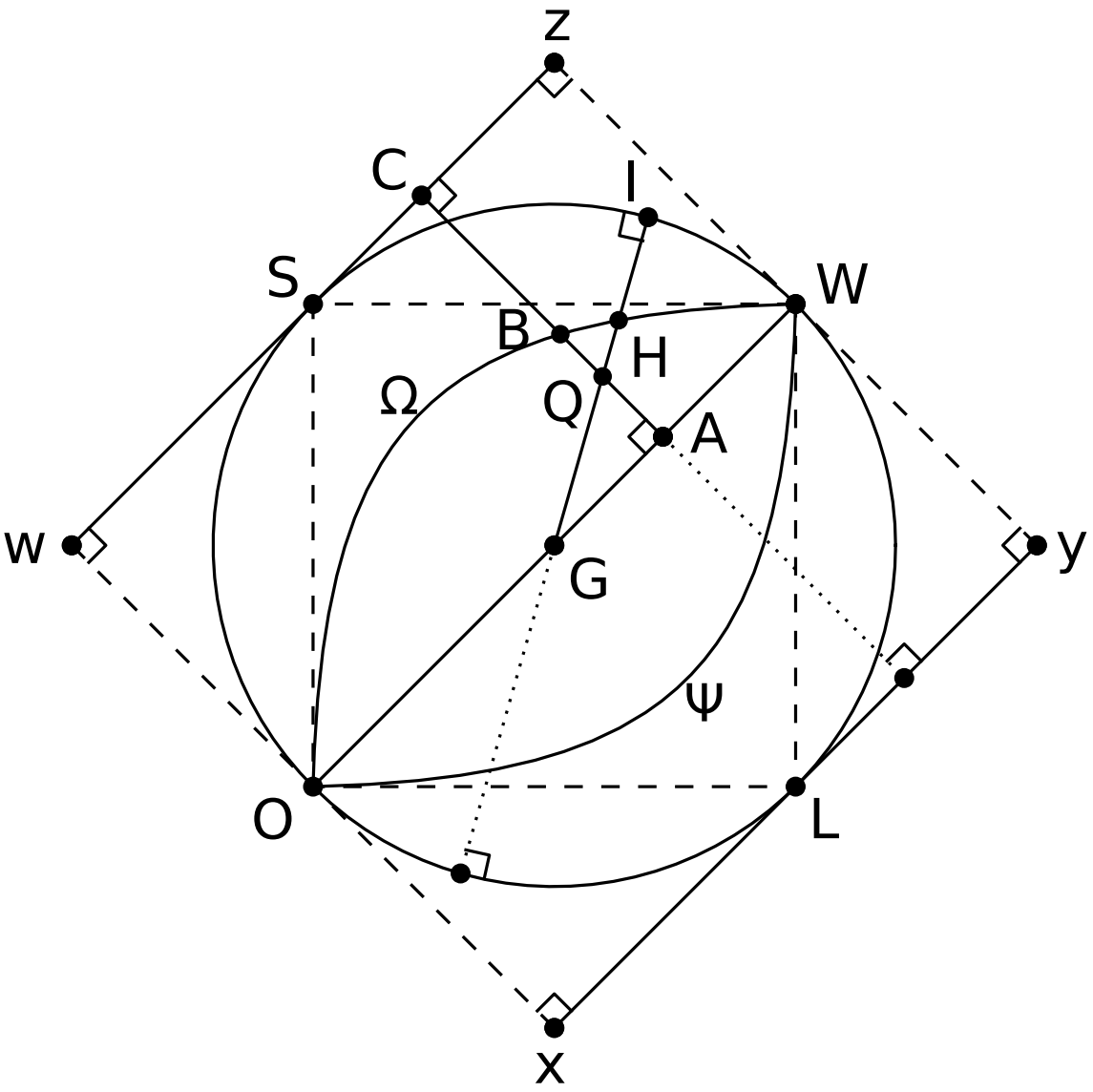

Figure 2.12: Schematic of the calculation of various quanifications of colourfulness for a quantum catch specified by Q for a dichromat ($n = 2$). See text body for details.

euclidean distance from the centre to Q. The corrected spherical colourfulness is calculated by finding the intersection of the spherical ray that passes through Q ($\overline{GI}$) with the colour solids boundaries. In this case it is the intersection of $\overline{GI}$ with Ω at H. To correct the colourfulness we simply divide by the length of $\overline{GH}$ to get $\bar{J}_R = \left|\overline{GI}\right| / \left|\overline{GH}\right|$. Similarly, we calculate the cylindrical colourfulness as the length of the line through Q, which also passes through the line $\overline{OW}$ perpendicularly. This happens at A, so the spherical colourfulness is given by the length of the line $\overline{AH}$ – the correction factor is $1/\left|\overline{AB}\right|$. The function $h(a, b)$ in the text is defined to find points lying on the colour solid. In terms of this function we calculate H as $h(\overrightarrow{GH}, \overrightarrow{OG})$ and B as $h(\overrightarrow{AH}, \overrightarrow{OA})$. Note: High dimensional cylinders are

classified by dimension of their capping spheres and the dimension of their sides, formally the unit a,b-cylinder is given by the Cartesian product $S^a \times [0,1]^b$ – a is the number of sphere like dimensions and b is the number of cube like dimensions.

2.5.2 Hue, Saturation and Lightness

Colourfulness (in both the cylindrical and spherical systems) can be naturally identified with the more established notion of 'saturation'. Saturation, when taken as a proxy for colourfulness, has an interpretation in terms of the number of spectra[12] that can produce a colour, and for similar reasons, the ability to predict the location of colours for one observer from another.

For value, we naturally have an axis lying between black and white, similar to the 'lightness' axis of RGB devices. Such an axis exists for all observers. The spectra given by $s(\lambda) = k, k \in [0,1]$ *always* lie along this line. However, when we are talking about the world an animal experiences, we would like to not have to refer to spectra at all. Probably the best interpretation of this axis is as the convex combinations of black and white.

For human colour vision this just leaves the hue. The remaining dimensions of colour solids have no easy generalisation. It seems it is just the remaining coordinate that 'falls out' when we identify lightness and saturation. The properties I have discussed so far, then, provide the basis for a naturalisation of the hue, saturation and value axes used in many human colour systems (they are defined e.g. CIE, 1931, 1976), although, as I have already indicated, this naturalisation is not complete until a behavioural role is explicated. I will do this in later chapters.

2.6 Summary

The discussion in this chapter has been about the gross geometry of colour spaces – which colours are next to others and which ones are not next others – which colours are possible and which are not. We know for example, that black or white objects will fall in exactly the same relation to other colours for all observers. We also know our ability to perceive relationships between other colours as seen by other animals is dependant on how colourful we perceive that colour to be. I am often asked about whether some bright coloured biological signal can be seen by other animals - clearly the answer is almost certainly yes. Nonetheless, it seems that people still consider it reasonable to think that

[12] or physical processes.

we need strong evidence to assert that a signal is obvious to another animal (Penney et al., 2012), but the chances are, if a signal is colourful to us it is colourful to the animal too.

Colour spaces display the relationship between different colours in geometric form and it is always the relationship between colours we care about. But, in this chapter I have not really discussed how we measure distances between colours. It is quite reasonable to say that I have only discussed which colours are allowed to exist given physical and physiological constraints. The involved subject of measuring distances between colours, from it's philosophical basis to application, will be covered in chapters 3, 4 and 5.

Part II

Colour and Measurement

Chapter 3

Measuring Living Things

"Man is the measure of all things"

Protagoras, 420 BC

The purpose of this chapter is to establish the philosophical grounding for the models which I develop later in this thesis. The argument I present is founded on the idea that to identify an organism to identify a agent - an subject that we consider justified in treating as animate: one for which beliefs and desires lead to *actions*. I will go on from this premise to discuss information theoretic measures in this context, with the ultimate goal of elucidating the foundations upon which the application of information theory to colour is, or at least should be, based.

The argument I am about to present is aimed at justifying the more formal theories in later chapters. Without a grounding in a more general theory, the decisions made in the derivation would be rather arbitrary. Importantly, the aim of this chapter is to lay the foundation for the theory of risk, and emphasise the importance of using relative measures (divergences) as opposed to 'absolute' measures such as entropy. This requires us to discuss the meaning of these terms, and for that, we must first discuss measurement.

Measurement is dependant on the methodology of measurement, and the choice of methodology is far from uniquely defined. Importantly, in probability theory, the set of possibilities, the ontology or the support, is a limitation on our objectivity as it can only be obtained by a judgement that can be at best held up to empirical observation, but never truly specified.

3.1 Perception, Behaviour and Purpose

Phrases like 'perceptually uniform' and 'perceptual space' are commonplace in psychophysical discourse, and the idea that a colour space is a perceptual model (in the sense of describing 'internal' appearance or psychological phenomena) is widespread. However, my concern here is not with perception, but behaviour – or more precisely, perception only to the extent that it is behaviourally observable within the constraints of a well defined experiment.

Though there are common limits to behaviour and to perception, the two are not the same. Perception contains objects and qualities which can never, in their totality, manifest as quantifiable behaviour. My perception of some detail may never have behaviourally observable consequences, at least not to the point of deduction, by another, of the entirety of that perception. I might be placed in a situation where I have a choice of doing one thing or another based on a judgement about something's appearance, but given the datum of my choice (or any data that one cares to measure), using it to reconstruct my perceptions over the duration of my judgement would be impossible. Similarly, there can be behaviours that are not fully part of perception. I may behave in some way that I am not aware of - tapping a foot or stroking my beard: I might perceive a consequence of this or that action, or deduce *post hoc* some aspect of it from my current perception, but this is not the same as being directly aware of it. In any case, even were it necessary for my action to enter my perception partially, it is certainly possible (if not necessary) for it not to in its totality.

This hints at how I wish to differentiate between perception and behaviour; the former is a first-person view, the latter a third-person view. Theories of perception concern what is *immediately present* to me or to others, but theories of behaviour are concerned with *how others appear* to me or to others.

A behavioural theory relies on our ability to identify agents and their actions, it does not tell us how to do so. Indeed, the field of ethology has proceeded without any great need to define what an organism is (Queller and Strassmann, 2009, for a recent attempt) – agents and their actions are apparent to us and it is from this 'given' that organisms are identified and behavioural theory is formed. A behavioural theory relies on perception and the 'standing out'[1] of organisms against the rest of the world. Whatever the criteria or mechanism for this judgement is deemed to be, the meaning of doing so is clear: By recognising something as an organism we are recognising that it acts with its own goals

[1] or *ek-stace* as used by Merleau-Ponty (1945, p100).

and with its own understanding of the world.

It is action for a purpose, be it teleological or teleonomical[2] in nature, that separates organisms from non-organisms and behaving from mere happening. A behavioural model then, treats organisms as agents with goals and beliefs, where actions are decided by the expectations of costs or benefits as shaped by an understanding of how outcomes may be achieved.

3.1.1 A Technical Note on Agency, Organisms and Genes

At this point I should clarify one aspect of my stance on the nature of the goals and beliefs of the organisms. For the purposes here, beliefs and desires need not be considered as more than things that we impose upon the world around us, be it voluntarily or otherwise. However, I am not agnostic towards the general use of agency. Far from it. When faced with even the simplest of cells, a physical-chemical approach becomes impossible as a living cell, like the ship of Theseus, is in a constant state of flux. The cell is not defined by atoms, or by chemical interactions, but by macroscopic properties; namely those that correspond to our judgement of it as a freely acting entity. Indeed, this principle lies in the origin of the word 'organism' itself: cognate with 'organ', it stems from the concept of a (usually) physical entity that communicates on behalf of another physical or non-physical entity (e.g. A pipe organ is the physical means by by which the non-physical component of music is communicated). As such, an organism is the physical counterpart, the "material structure", of a biological agency (OED, 2012). It is also worth noting that none of these words are cognate with 'order', despite the similarities of their modern interpretation (*ibid.*).

It is all well and good to say that what I'm describing here is a concept of the organism that follows the spirit of the original term, but this is no argument for its correctness. For now, I will simply say that the description I have given is sufficient to define an organism for the purposes of this thesis, and, that I will be happy to resign my argument when or if someone provides me with a physical, mechanistic description of an organism.

Organisms as described here share some common ground with those of Dawkins (1976, 1982) in the sense that when we see an organism, we are seeing it as, or resulting from, goal directed action. For Dawkins, the only viable goal is maximising inclusive fitness as described by natural selection. I approach this more ecumenically, the goal may be whatever one chooses as long as it can be justified. A gene centred justification can be perfectly

[2]Teleology is purpose in reality, teleonomy is 'as if' for a purpose - see Bedau (1991); Weber and Varela (2002) for a pro-teleological accounts, or Reese (1994) for the contrary

fine, but so can those justifications that do not mention genes at all. Fundamentally, the difference lies in an approach to science, a strict gene centred approach suggests that one can and should eliminate this kind of judgement from scientific enquiry, where as I think that such an endeavour is impossible.

3.2 Introducing the Formal Tools

Now I shall proceed by describing the formal tools that I am using, as well as providing justification for using them. To mathematicians this may seem to be an overly verbose description of a fairly simple formal theory that would be found in a sufficiently detailed book on probability theory. For those more verbally inclined it may seem like long and overly formal description of some common sense principles. However, seen as intended it is a description of the correspondence between the mathematical formalisms of probability theory and some very general concepts that underlie all of scientific investigation. This understanding is necessary if we are to be clear about exactly what is measured and predicted.

3.2.1 Identity and Equivalence

The notions of identity and equivalence a fundamental to probabilist and informational measurements, as these will eventually become part of the theory, these foundational notion requires some attention. The famous Gibbs paradox highlights how calculation thermodynamic entropy is dependent on what we consider to be the same or different (in thermodynamic terms, indistinguishable) a problem which is far more profound when the idea of entropy is taken away from its original context. Furthermore, concepts pertaining sameness and difference are fundamental to the concept of number (Husserl, 1891) and thus, are present in the very tools foundation of measurement. The formal languages of mathematics are abstractions from based upon these ideas (*ibid.*). I shall start with the concept of identity - the idea that two things are *exactly* the same.

It is tempting to think that it is possible for two objects to be identical, at least in theory, for it happens in mathematics all the time. For example, one might consider two hypothetical apples, both internally identical so that for every molecule in one there is a corresponding molecule in the other. The two are chemically identical; each constituent molecule them standing in the same relation with every other constituent molecule. Thus, one may wish claim that there is no difference between the two. Of course you would be right if position and orientation in space were ignored. But being able to reference

their spatial location relative to each other allows us to distinguish them, indeed, *it does* distinguish them. If it was impossible to do even this, then one could not reasonably claim that the apples were different things, nor could one claim that there were two of them in the first place. This highlights the difference between equivalence and identity: equivalence is the claim that there are two different things but we are unable to distinguish them by some or other measurement, and identity is the claim that actually there is just one thing. In the case of the two apples, they are equivalent if we ignore spatial difference *and are aware that we are ignoring it*, they are equal if *either* they exist in the same position space, *or*, if we ignore, consciously or otherwise, the *possibility* that they *may* differ in their location.

Let us make this more concrete with an example from the history of science. Before the 1820s, it was thought that all molecules could be completely described by the number and type of atoms within them, chemical species were identical when they had the same constitution as determined by quantitative techniques such as titration[3] and gravimetry[4]. In the 1820s a German chemist named Friedrich Wöhler synthesised a compound known as silver cyanate, on analysis of this he discovered it had the same formula as another molecule, silver fulminate, which had been synthesised by another German chemist, Justus Liebig. As the understanding at this time was that species with the same chemical composition were equal, so, when it was observed that these two substances behaved differently (unlike silver cyanate, silver fulminate explodes) there was confusion and hostility between the two chemists (Esteban et al., 2008). The situation was resolved with Jöns Jacob Berzelius' concept of isomers (literally *same-parts-ers*), in which he suggested that the atoms could be joined up in different orders (*ibid.*). Through this development, what was once an identity became an equivalence - conceptually, every molecular species differentiated into the permutations of its atoms, the unity of the chemical composition became manifold.

This happens again and again in science. With new measurements comes new differences, and it is not surprising that isomers were not the end of the story. Now, in the days of structural biology, it is not even the order that the atoms within the molecule that are of greatest concern, but their overall shape. But of course, this is just a newer way for something to be different, an there are yet new ways of finding finer differences still.

All I wish you to take away from this is that equivalence is just an identity that has been shown to be false, and that in all likelihood, equivalence is the fate of all equalities.

[3] Analysis of volumes and concentrations.

[4] Analysis of mass.

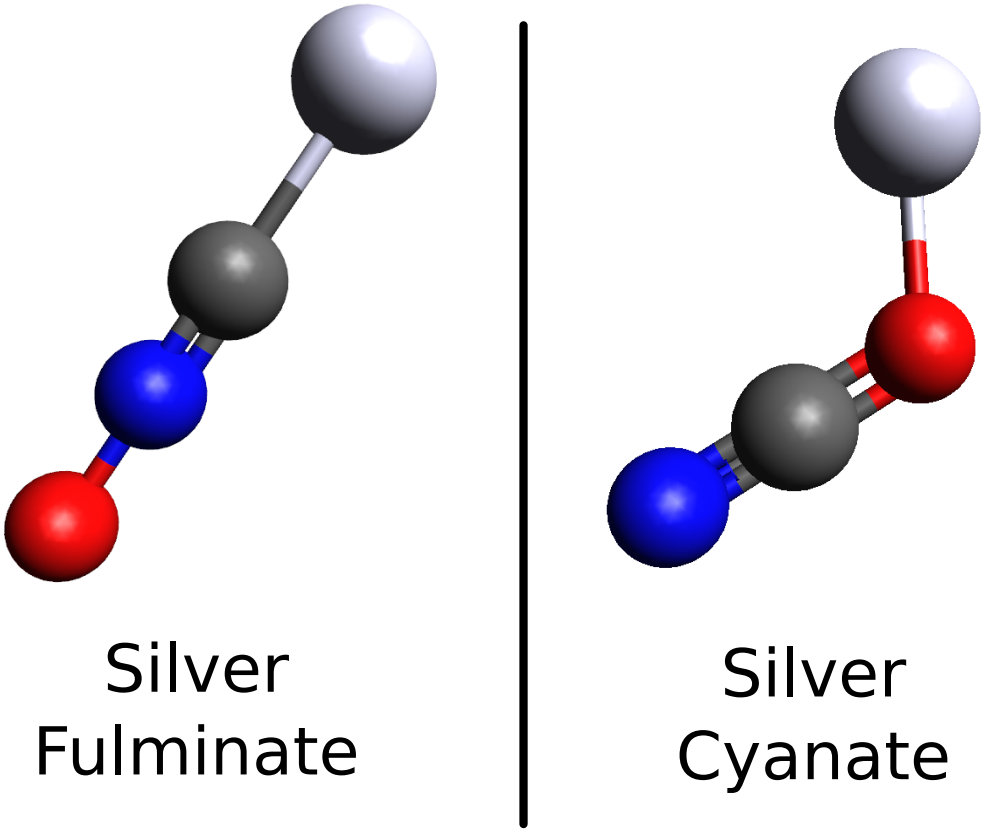

Figure 3.1: No longer equal, but compositionally equivalent. Left: Silver fulminate (AgCNO). Right: Silver cyanate (AgOCN). Red – oxygen, Blue – nitrogen, Grey – carbon, Silver – silver. The molecular bond lengths and angles are calculated with a molecular dynamics energy minimisation (universal force field) and may not completely correlate with those determined crystalographically or using more detailed simulations

3.2.2 Pragmatism

But we cannot work in a world where everything is differentiated into an infinity of possibilities, nor do we want to. Even knowing that a coiled spring can be described in terms of subatomic particles doesn't mean that this is *the* correct way of thinking about it. And it's certainly not the most practical mode of description: it wouldn't help me design a car's suspension system! For this, its possible that Hooke's law will suffice, failing that a non-linear model, or continuum mechanics, etc. Practical application requires parsimony, but moreover, it's quite possible that parsimony is as good as it will ever get.

If we do not (or cannot) choose to describe the infinite variety of the world, at what point should we draw the line with our finite one. Which formalism should we choose?

Should it come from principles, or from intuition?

Indeed, what I am expressing is a central thesis in pragmatism, often stated as *Peirce's Pragmatic Maxim* (Peirce, 1878):

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.

Not only are we required to make assumptions about the latent composition of things to make sense of our measurements (Lewis, 1952) but we should embrace, or at least be conscious of, this need when approaching study of them.

Any quantitative theory has at its core a set of atomic components that are not identical by definition. In the example above, before the theory of isomers chemists worked with a set of components that could be enumerated as ratios of elements, each of these element ratios were not only bound by equivalence but identity – at the time they were considered to be all there was to know about the composition of a chemical species. Any chemical species could be represented by a member of the set $\{XY_aZ_b\ldots; a, b\ldots \in \mathbb{Q}_{0/+}\}$ where each $X, Y, Z\ldots$ was some known or unknown element. At that time, if I synthesised some chemical, it was assumed to be *exactly one* of these. This kind of set – a set that contains all the different allowable possibilities for something to be – I shall call 'the ontology' for now. Later on I shall give it the more conventional name of 'the support'.[5] It is the enumeration of the ways of being (Greek: ὄντος, ontos). In physics, in particular statistical mechanics, a way something can be is called a 'state'. Using my wording, in the example above we would say: after discovering the possibility of isomers, a new ontology emerged – one that accounted for the new observed differences.

[5] It seems that this term has some root in the philosophy of Leibniz, where the support is similar to the Aristotelian concept of substance: something that exists and has properties (Broackes, 2006).

Formalisation

The formalisation begins with a set of atomic possibilities $\mathcal{X}$, the support, or ontology. This set may be finite or infinite[a]. Each element of the set is mutually exclusive, or, in other words, different and distinct.

We also require a way of addressing these possibilities. To do this, we introduce coordinates and/or indexing. I will use x^i to denote a continuous coordinate belonging to a coordinate system $X = (x^1 \dots x^n)$, so that each point in the coordinate system corresponds to a different mutually exclusive possibility. If there are discrete possibilities I will just use labels, although I will use indices such as i, j or k when it is efficient to do so (such as describing a set of N discrete possibilities by $\{x_i \mid i \in \{1 \dots N\}\}$).

[a]But it must be measurable – more on this later.

The Ontology and Measurement

The question arises of why we should need to decide on an ontology at all, especially given the vanishing chance that it is absolutely correct.

In a sense we do not. We can always acknowledge the fact that our picture is incomplete and doing so does not by necessity affect our understanding. However, it does affect one thing that is of utmost importance to scientific understanding: *quantification*. I will demonstrate this by comparing two chemical theories, one is very well established and practical, the other is designed for this explanation and of little practical value beyond it. Later, we will see this phenomenon is a stumbling block in the measurement of entropy.

The 19^{th} century chemistry was a time before the widespread acceptance of the concept of *atomic mass*. Often, all that was considered was the ratios of elements. For example, ethene and cyclobutane have the same elemental ratio, two hydrogen atoms for each carbon – written using the *empirical formula* CH_2. These days, we would describe these two chemicals with different *molecular formulae*, C_2H_4 and C_4H_8 respectively (or even represent them schematically). We would know that in one gram of ethene there is twice as many molecules than in one gram of cyclobutane. Once we establish a constant (Avogadro's constant, N_A) and choose weights for each type of atom, we can put an actual number on it[6]- there is approximately 2×10^{22} molecules in a gram of ethene and 1×10^{22}

[6]The calculation is $N = \frac{mN_A}{n_H m_H + n_C + m_C}$, where N is the number of atoms in mass m, N_A is a constant equal to 6.022×10^{23}, n_H and n_C are respectively the number of hydrogen or carbon atoms in a molecule

in a gram of cyclobutane.

But let us imagine that we didn't know how to do this and simply thought the empirical formula was sufficient. This doesn't prevent us from making some kind of molecular theory. Let us do this now. Instead of molecules, lets say that a mass of chemical is composed of something else: polecules. Each polecule has an integer numbers of each type of component atom, but unlike molecules, polecules contain the only minimal number of atoms required to achieve this, thus C_2H_4 and C_4H_8 refer to the same type of polecule: CH_2. We can say how many polecules are are in a given mass of a given chemical, just as we can with molecules. There is some physical significance to polecules too, we could identify individual polecules within a mass to the same extent that we can molecules. Ethene is a pair of polecules and cyclobutane a quartet. The molecular and polecular descriptions are both valid within certain constraints; in a world where only the relative proportion of elements matters there is no difference between them at all. Outside of this domain of the experimental techniques of pre-19$^{\text{th}}$ century chemists the two are of course very different[7], indeed, despite a physical correspondence, polecules have proven far less useful, far less illuminating, providing far less resolution of the world the attempt to explain.

More formally, every molecular formula of the form $X_nY_{an}Z_{bn}\ldots$ (for integer n) is mapped to an empirical (polecular) formula of the form $XY_aZ_b\ldots$. Effectively, we are still measuring the same physical thing, only we are choosing to multiply the empirical formula by n. The mapping from molecules to polecules is many-to-one, reflected in the limited applicability of empirical formulae when compared to molecular formula. These two measures of a pure substance are exactly the same except for a difference in the ontology, to find the number of molecules/polecules we perform exactly the same procedure. Firstly we find relative mass corresponding to the formula, then we divide our mass by this number and multiply it by Avogadro's constant. There is no difference in *how* we measure, nor in the *physical substance* that we are measuring, the only difference to be found is in *what we are measuring it in terms of.*

In this case it is fairly obvious the polecules and molecules are different things, but as we will see, such differences become obfuscated in the language of information theory. The difference between molecules and polecules is easy to keep in mind – they have different names – but in information theory, everything is measured in *bits*.[8] The word *bit* hides a

and m_H and m_C the relative atomic masses of one of those atoms (a mass per N_A atoms quantity).

[7]They may be different, but they can be related to each other: the number of polecules in a pure substance is always a strictly positive integer multiple of the number of molecules.

[8]Bits, or multiples thereof: e.g. nats (or decibels Jaynes, 1994).

gross difference between informational measures and provides a false sense of generality.

To observe consequences of our choice of ontology, we must show how it results in a measurement that can correspond to some real world phenomenon. In the hypothetical case I just described, the physical meaning of either molecules or polecules requires adding a way of getting a measurement from the ontology – in this case in the form of amount of molecules/polecules. We could then go on to test this against some other, physical, measurement of the number of molecules/polecules, such as reactions that only proceed on a single molecule. If we had an ontology that corresponded to the collection of points found along the edge of a ruler we would need to define how to take one of those points and turn it into a distance. Informally, a sensible way of doing so might be 'counting' how many points there were between that one and which ever one we call zero.

So, not only do we need to state those things that we are considering, but also we need to state how those things should determine or relate to the measured, observable world. Before we define a way of measuring things our ontology is intangible – we merely have a collection of things without meaning. The (abstract) measurement of the members of the ontology enables (physical) measurement and defines the mode by which any ontology is validated or invalidated.

Formalisation

Sigma Algebras. Before talking about belief and knowledge, I will quickly explain the established underlying formalism. The use of σ-algebras allows us to talk about events – not just the different elements of the ontology, but combinations that might not be mutually exclusive.

Measure theory begins with the ontology (support) and on top of adds a list of the combinations of those things about which a measurement can be made, the measurement may be something tangible, like a volume, or something else, like a probability. This is called the σ-algebra, Σ (see footnote[a]). We generally require the sigma algebra to reflect valid statements in a reduced classical logic (again, see footnote) – one where we do not have implication as such. Σ contains only subsets of $\mathcal{X}$, $\forall E \in \Sigma : E \subseteq \mathcal{X}$. Then we define, when it is possible, a function μ that assigns a positive real number: $\mu : \Sigma \to \mathbb{R}_{0/+}$. This happens quite naturally (but not uniquely) for the ruler example, in such a case, μ is known as the Lebesgue measure, and corresponds to our usual notions of distance (as well as area and volume in higher dimensions). This kind of formulation will be needed if we wish to consider ontologies which are described using continuous variables.

In the next section I will describe how beliefs are usually formalized in this framework.

[a] Often we can assume that the σ-algebra is the power set of $\mathcal{X}$: $\Sigma = 2^{\mathcal{X}}$, but this is not true by necessity. Allowable σ algebras are given by the axioms: (a) Non-emptiness: $\Sigma \neq \varnothing$, (b) Closure under (countable) union: $E_1, E_2 ... E_n \in \Sigma \implies \left(\bigcup_{i=1}^{i=n} E_i\right) \in \Sigma$ (c) Closure under complement: $E \in \Sigma \implies (\mathcal{X} \setminus E) \in \Sigma$. In probability theory, these correspond (approximately) to: (a) A requirement for at least one possibility and it's negation. (b) The necessary validity of asking 'p OR q' for any valid p and q. (c) The necessary validity of 'NOT p' for any valid p

3.2.3 Belief and Knowledge

Knowledge endows its possessor with the ability to act with finesse. For this reason it is a tangible subject of scientific enquiry. One would be hard pushed to say that given the observation of someone riding a bike, that they do not *know* how to ride a bike. But, we cannot talk about knowledge without talking about the related concept of belief and drawing a distinction between the two.

One of the earliest distinctions was made by Plato (Jowett, 1871), often translated as

"knowledge is true belief".[9] Explicating the general meaning of 'true' is far too difficult an exercise for the purposes here, so I shall not use Plato's statement "as is". Nonetheless, it's a good starting point – knowledge is indeed a type of belief that can be shown to be 'correct' according some chosen criteria. I shall make no claims as to whether one should consider this correctness as truth in any sense.

The concept of an agent provides us with a good platform for investigating this. If we are consider something as an agent on the basis of rational action *towards* – or *as if towards*[10] – some goal, then we already have a good candidate for what could be taken to be meant by truth in Plato's canon. Knowledge can be measured by the ability to achieve a goal, as long as that goal is desirable; On being given a desired goal we are able to judge the efficacy of a belief to that end, and in doing so we are treating that belief as knowledge about achieving it. Plato's canon should perhaps be *"knowledge is useful belief"*.

Without the assumption of some goal, belief would be without behavioural consequence and it would be impossible to quantify it. For example, one would not trust someone to be honest if it was clear that they had either motivation not to be, or lacked such motivation entirely. Effective communication of belief requires the person being informed of it to recognise in the one informing them a motivation that would reveal their true belief. It is probably worth noting that it is exceedingly rare for us to consider any organism as having a complete lack of motivation towards some goal or other – such a lack is incongruous with our intuitions about agency.

Given what I have said so far, we might consider modifying Plato's cannon further to refer to *measurable* belief.

This all presupposes that we may correctly judge what is and what is not a goal that the organism desires, a difficulty that all such arguments come up against. This is well known in biology as it proves to be a major difficulty when discussing adaption (Gould and Lewontin, 1979) but it is a necessity that stems from our decision to consider agents as the subject of enquiry (Weber and Varela, 2002). Some goals are more reasonable than others, for example, staying alive[11] is in most cases fairly easy to justify, but the further away from fundamental needs one moves the more difficult such justification becomes. No doubt this is reflected by the many varied schools of psychoanalysis. On the other hand, if we choose to throw out the purposefulness of behaviour we should also throw out the

[9] Generally it is now stated as "knowledge is justified true belief", this is completely consistent with what I will say in the following pages as they describe a way in which a belief can be justified.

[10] If one chooses to acknowledge this distinction.

[11] Compare with 'replicating' (Dawkins, 1982).

concepts of agents and organisms and with them biology, social science, economics and the entirety of the arts.

To develop our understanding of belief we must first consider the ontology about which any belief is held. This is probably the most difficult part, as it is very easy to conflate an agent's ontology with our own. There are two ways in which one can make this mistake.

First, there is the obvious mistake of assuming that some object that we understand to exist also exists for some other agent. For example, thinking that cups of coffee, as such, exist in the world of an ant.

The second way is far more surreptitious and is only usually realised when we recognise that we have ended up asserting that *our knowledge* of an organism is accessible to that organism itself, or in the extreme, that it has direct access to its own physiological state in its entirety. It is obvious that a person does not have a direct awareness of, or present behavior that can illuminate, the entire state of neuronal potentials or concentrations of neuromodulatory molecules, but avoiding such implications proves difficult if we are going to talk about organisms both as agents and as physical entities, in part comprised of neurons and neuromodulatory molecules. If we cannot say that an organism makes decisions directly on the basis of its physiological state we must instead say how we can use physiology to inform our understanding of goal directed action. If physiological and behavioural descriptions are not *automatically* compatible, and we wish to integrate them, we must say how we can make them compatible.

In the proceeding chapters I will discuss the limitations upon action as informed by physiology. For example, it is very useful to consider the limitations of discrimination caused by the indeterminacy of the state of photoreceptor cells (Buchsbaum and Goldstein, 1979a,b; Vorobyev and Osorio, 1998). As the state of a photoreceptor determines the ability to perform particular tasks it is very tempting to say that an organism holds some kind of belief about that state. There is a problem with this, but it is not that the organism does not hold such beliefs! It may or may not. We cannot know beyond what we choose to test. The problem is instead that, as third person observers of the organism, the interrogation of these beliefs requires that we measure them against some goal. Unless the goal the organism desires is to report the exact degree of photoreceptor excitation it would be an error to take it to be the object of *the organisms* belief[12].

[12]We should none the less remember that this does not directly exclude the *possibility* of beliefs being held about it

Ratifying Behaviour with Physiology

So how then do we ratify the behavioral with physiological? I shall approach this by explaining how we do it in practice and argue why this approach is justified.

The first step (of three) is to only consider behaviours in such a way that allows a comprehensive description: we select aspects of behaviour to be the only things that matter for our purpose - choosing a collection of behaviours which are reasonably assumed to be complete. An example of doing this is in two alternative forced choice paradigm (2AFC), where the organism is only considered as doing one of two things. In this case, it is not that the experimenter is unaware that the organism behaves with infinite subtlety, but they are choosing to look at behaviour in a way so that every action of the organism during a trial *must* belong to one of two classes.[13]

We can phrase this step in another way: As scientists we wish to treat the system as causally determined, but for some reason we have chosen to study something that we consider not to be – something with freedom to act on its own behalf – something *autonomous*. To make this thing something that we can study scientifically we must rid the thing of it's autonomy.[14] We do this in two ways as, firstly by putting it in a situation where its actions are physically constrained, i.e. put it in a laboratory setting, secondly, we course grain the behaviour; choosing to look at it as something with no other actions than the ones we select. In effect we redefine the rat in a maze so that it is identical to a 'left-right turning machine' or a 'getting-lost-not-getting-lost machine'.

Of course, as I have indicated before, the validity doing this rests on a judgement that the constrained behaviour in the lab reflects the non-constrained behaviour in the world at large.

The second step is to control the goals of the agent by either providing sufficient motivation towards a goal we choose, or by identifying a goal that the organism already has. I have already discussed the relationship between goals and beliefs above, but I should once again say that we must make a judgement about whether we can rightly claim that we know the goal of the organism. We must coerce it.

The third and last step is the one which has important consequences for information

[13] I should probably point out that when we generalise to situations outside of such paradigms, we are making the assumption that the goals of the organism in the new situation are suitably aligned with the goals of the organism in the experimental paradigm.

[14] Put like this it is hard to ignore the potential for ethical implications of performing even non-invasive behavioural experiments.

theoretic interpretations. This step is the constraint of the ontology. Concisely, we must choose a collection of physiological properties that we take to detail every relevant physical parameter. We can write this another way:

Assumption: *our ontology is sufficient to resolve the beliefs of the organism to the degree that is observable using our class of behaviours.*

This can be seen as simply an application of the the pragmatic maxim.

Under this assumption there are two ways that the ontology of organisms belief may differ from ours as observers that have no consequence: (a) The organism's ontology may be 'coarser' than ours; there is a many to one mapping from the states of affairs in our ontology to that of the organism. In this case we would discover that there are aspects of our ontology of no consequence for behaviour and we do not need to do anything, except perhaps reduce our ontology for the sake of parsimony. (b) The subject may hold beliefs about extraneous states of affairs; about things which we (the experimenters) neither represent in our ontology nor play a role in determination of the experimental result. This clearly does not matter.

The making of the above assumption is further justified as it leads to testable consequences in the sense that it is possible to demonstrate that it was used incorrectly. We can do this by simply extending our ontology and looking at whether it allows us to better predict behaviour. In other words, we can consider more and more factors, and look in more and more detail, until we have confidence that we have sufficiently resolved the states of affairs that lead to a particular behaviour. The limit, where for a particular behaviour we have assured or assumed that our ontology is sufficiently detailed so that our ability to predict the behaviour bottoms out, is the *psychophysical limit*.

It is not just the ontology that we must be careful about projecting onto the organism, but our beliefs too. In fact, taking a subjective view of probabilities as we must to discuss beliefs, it is not obvious how we may talk about *another* agents beliefs at all. Indeed, we cannot if we take "agents beliefs" in the sense of some objective fact that we are determining. We must remember that the beliefs *we consider the agent to have* relate to the goals and actions *we consider it to have*. To treat this correctly, we must take the beliefs of the agent to be anything sufficient action towards the goals with the degree of skill that we observe. We should then, of course, measure them in a way that corresponds to this.

By definition, in specifying a non-deterministic relationship between a physical and another (probably physiological) physical property we are saying that no-one who has access to only one can have complete knowledge of the other without additional data

about it. Thus, when we consider an organisms physiological state that is is specified in this way we are saying that the organism cannot have complete knowledge of some physical property. This means, that although we are not forbidden to make claims about what an organism *cannot* know, they must be justified by a argument for why it is reasonable to assume the organism cannot obtain data that allows it to know more.

A simple example of such a justification is this. I put a black ball and a white ball into an opaque bag and shake it around. The colour of the next ball I pull out will be underdetermined for me. It is then easy to justify that it should be underdetermined for others too: unless they are in a position to see into the bag and judge which ball I will lay my hand on first, or track where each ball lies within it, they have the same information that I do about which ball I will draw. Their knowledge, as judged against their accuracy of prediction, will be the same as mine.

The same applies to cases such as the modelling of photoreceptors, in this case a justification can be made based on the greater investigatory power that scientific instrumentation provides us with.

To reiterate what I have said:

We know what the organism wants to do and we know what it has done, and, although we *don't know* all the details of the process by which it has chosen a given action we know all the data that it could have used to make that decision.

Summary of this Section

What I have said so far has an interpretation that I have been avoiding up to now, but one that is central to this thesis. The interpretation is information theoretic, and what I have said above can be re-expressed as:

1. Make the organisms behaviour well defined so that it can be measured.
2. Make the physiological parameters well defined so that they can be measured.
3. Define a goal so that we have a clear understanding of what information is required to achieve it i.e. the information required for particular beliefs to be rationally held.
4. Investigate the physiological limitations of achieving that goal i.e. the information present to the organism from the physiology.

Expressed in this form, I hope that it is fairly clear. Once the formalisations are complete I shall explain why this is relevent to information theory and perceptual theory more generally.

Formalisation

Belief can be adequately formalised by probability theory. Whilst there are other formalisations (e.g. fuzzy logic as popularised by Zadeh, 1968), probability theory has the benefit of fitting with classical conceptions of rational thought. I shall not argue this here, as the argument has been made by many times before (Cox, 1946, 1961; Jaynes, 1994). Here I review the widely accepted formalisation of Kolmogorov (1956), which differs to the foundations favoured by some subjective probabilists, but it is still the most popular and fits with the rest of the formalism here.

Probability is a particular kind of measure, so, we have already laid down much of the groundwork. Indeed, we already have a space of things to be measured $\mathcal{X}$ and a list of the combinations of which we can measure Σ. Probability is then formalised by defining a function μ that takes some measurable combination of things from Σ and assigns it a number in $[0, 1]$. So $\mu : \Sigma \to [0, 1]$ with the constraint that the probability of the system not being in any state is zero $\mu(\varnothing) = 0$, the probability that it is in *some* state is one $\mu(\mathcal{X}) = 1$ and that for any disjoint subset of Σ: $\{E_i\}$ that $\mu\left(\bigcup_i E_i\right) = \sum_i \mu\left(E_i\right)$. This last requirement is a refined version of requiring that the measure of a set from the σ-algebra, E, is the sum of the measure of its elements, $\mu\left(E\right) = \sum_{x_i \in E} \mu(\{x_i\})$. We are required to use a more nuanced definition than this as it does not work when the measure of the individual elements, $\mu(\{x_i\})$, is vanishingly small (as it is when $\mathcal{X}$ is used to represent a continuous variable).

If we assign logical statements to the elements of Σ then we can find corresponding probability statements. For example the probability of A and B, $\Pr(a, b)$ corresponds to $\mu(A \cap B)$; not A, $\Pr(\neg a)$, to $\mu(\mathcal{X} \setminus A)$; A or B, $\Pr(a \vee b)$ to $\mu(A \cup B)$; *etc.*

Conditionals are slightly more difficult so I shall ignore some details here, specifically the problems concerning conditioning on set x where $\mu(x) = 0$ (see e.g. Jaynes, 1994, chap 15.7). Essentially though, we simply define the probability of A given B, $\Pr(a \,|\, b)$ by using the law of conditional probabilities ($\Pr(a \,|\, b) \Pr(b) = \Pr(a, b)$) leading to a correspondence of $\Pr(a \,|\, b)$ with $\mu(A \cap B)/\mu(B)$. Difficulties associated with $\mu(B)$ being zero can be avoided by assuring that in such cases we are explicitly take some well specified limit (*ibid.*).

Now that we have a formalisation of belief. We only need to show the formalism of their connection to behaviour. This is done using the apparatus of risk. To apply it, we take it to be true that the organism has the knowledge of how to rationally act in a contingent manner upon their beliefs about the world. The type of rationality to which I refer is began with the work of von Neumann and Morgenstern (1944) but is now best identified with the notion of generalised rational expectations. In this case, formally, the problem becomes a general maximisation problem. We expect that an agent uses some kind of rule to compare the relative benefits of the various contingent behaviours. This is known as a decision procedure.

The formal decision procedure is the expression of a goal as an objective (target) function conditioned on the belief of the agent. This is then maximised according to some criteria (which can be also thought of as a part of the goal).

The assumption of a well defined goal is nescessary, and can been see as a problem for this theory. I will discuss this more fully in the later chapters, but there is one sense that a poorly defined goal can enter the formalism – by having a risk that depends on certainty. This can be thought of a way of catching the problems associated with ill defined goals – and as I will discuss in the next chapter, the seed of information theoretic measures.

Formalisation

Risk: Most generally, we say that a behaviour b from a set of behaviours $\mathcal{B} = \{b_i\}$ is assigned a value the corresponds to how well a goal is achieved, v_b. The behaviour that is enacted is then the one which has the largest v_b. There are many ways of assigning it, generally, it will involve (a) an ontology enumerating the contingent factors in making the decision $\mathcal{X}$ (and Σ) (b) a probability representing a belief about which state of affairs is correct, $\mu(x)$, (c) a function that assigns a the benefit to each one $-\mathcal{R}_b(x)$ (d) A function that dictates how these are integrated to judge each behaviour f. So we have $v_b = f(\mathcal{X}, \mu, \mathcal{R}_b)$. Often we can can write this as $\int_{\mathcal{X}} \mathcal{S}_b(x) \mathrm{d}\mu(x)$ for some $\mathcal{S}_b$. In the case of the Bayes risk based decision $\mathcal{S} = \mathcal{R}$. In the case where the decision is based on minimising the worst possible outcome (as measured by $\mathcal{R}$, known as minimax) we have $v_b = max_x \mathcal{R}_b(x)$, in which case we cannot write it in integral form. However in this case we observe that if we assume for simplicity that $\mathcal{R}$ is positive then $\int_{\mathcal{X}} \mathcal{R}_a(x)^n - \mathcal{R}_b(x)^n \, \mathrm{d}\mu(x)$ diverges to positive infinity if the largest risk belongs to behaviour a and negative infinity if it belongs to behaviour b, allowing us to rank the behaviours in the same way as the non-integral form. Although, the integral form is very general and many choices of f can be written as an integral of some $\mathcal{S}$ so that they give the same ranking of behaviours, it is not completely general.

3.3 Measures of Information

Now we are in a position to see how this relates to information theory. As the first part of this, I would like to identify four different types of measurement involved in information theory. It is good to make this explicit as it will allow me to both explain the motivation behind the approach used in later chapters and how information theory fits with what I have described so far. I will attempt to go through them in a natural way.

3.3.1 Geometric Measures

The class of measurements that I call geometric measurements are the kind of measurement that we make with rules, scales and stopwatches. They are when we measure the distance between two points in space or time, or the mass or volume of an object, or even the

fractal dimension of a coastline (Mandelbrot, 1982). They concern the 'world out there' (the word geometry, after all, stems from the Greek meaning 'measurement of the land', OED 2012).

Geometric measurements require method. For example, if I draw a number of different right angled triangles on a page and measured them using a standard ruler, I would find that the square of my measurements of the two shorter sides was equal to the square of my measurement of the longest side. If instead, I use a slide rule to measure exactly the same triangles, I will not observe the Pythagorean relationship between the squares of the numbers I read off.[15]

Geometric measurement has a normative character, there is no reason to choose a linear rule over the logarithmic slide rule other than convention and the ease of applying certain laws. The apparent arbitrariness of geometric axioms concerned many great thinkers of the 19$^{\text{th}}$ and 20$^{\text{th}}$ century, including von Helmholtz (1876) and the logician Alfred Tarski (Tarski and Woodger, 1938, paper II)[16]. This is before we even consider field of *mathematics proper*.

To make measurements of things in the world, we need a method - an agreed protocol by which we measurements are taken - so that we know exactly what it is we are measuring which rules apply to the measurements that we make. It is in such protocols, and their interpretation, that we find the role of judgement calls.[17]

3.3.2 Entropic Measures

Entropy is the earliest information measure. It was originally developed as part of Boltzmann-Gibbs statistical physics to describe thermodynamic entropy. It is usually expressed, for discrete supports (ontologies) as:

$$H = -\sum_i p_i \log p_i \tag{3.3.1}$$

[15]Say we have a right angled triangle with sides a, b and c, and we measure it with a normal ruler making measurements, R_a, R_b and R_c then we would observe that $R_a^2 + R_b^2 = R_c^2$, however, if we used a slide rule, getting measurements S_a, S_b and S_c, we would instead see the rule $e^{2S_a} + e^{2S_b} = e^{2S_c}$.

[16]Tarski was, like many logicians, about what was part of a particular formal system, and what we impose upon it.

[17]In the case of the Pythagorean relation, I expect it is so ingrained most people would make the same judgement about how to measure the lines to which it applies - but it is quite possible they might make mistake of judging it to be applicable where it is not, on the surface of a sphere, for example.

An Example of Bit Accounting

The appeal of the application of information theory comes from examples such as those found in Jaynes (1994). In certain cases it is possible to use information measures to say what number of statements about a state of affairs is enough to deduce exactly what is going on. This relies on the entropy being zero in complete certainty and maximal in complete ignorance.

Lets take the example of two people playing a very simple game with two different fair coins - one £1 coin and one £2 coin. The first person (Alice) tosses both of them keeping them hidden from the second person (Bob). We can measure what Bob knows about these coins by the entropy - there are four possible states: $(£1, £2)$ is one of $(H,H),(H,T),(T,H)$ and (T,T). The (maximum[18]) entropy, in bits, is then given by assigning the probability of $\frac{1}{4}$ to all states, so $-\sum p \log_2 p = -4\frac{1}{4}\log_2 \frac{1}{4} = \log_2 4 = 2bits$. Doing this gives us an upper bound on the ignorance of Bob. Being able to do this is essential.

Now, Bob is allowed to ask yes or no questions to Alice so as to determine the state of the two coins. There are a large number of questions he could possibly ask her, some of which are better than others. For example, he could ask "does the £1 coin show a head?", this would determine exactly the state of the £1 coin, meaning that the state of affairs, $(£1, £2)$, is given by (H,H) or (H,T) if the answer is yes or (T,H) or (T,T) if the answer is no. This question reduces the number of possible states to two, the expected change in entropy is the mean of $-2\frac{1}{2}\log_2 \frac{1}{2} = 1bit$ (answer is yes) and $-\frac{1}{2}\log_2 \frac{1}{2} = 1bit$ (answer is no), which, of course, is $1bit$. Bob's entropy with respect to the coins changes from $2bits$ to $1bit$. This binary question gives him $1bit$ of information.

If however the question was "does at least one coin show a head?" the possible states that the coins could be in are: $(H,H),(H,T)$ and (T,H) if the answer is yes, or (T,T) if the answer is no. The first outcome is expected to happen $\frac{3}{4}$ of the time, the second $\frac{1}{4}$ of the time. So the expected entropy of Bob's knowledge after getting an answer to this question is $\frac{3}{4}(-3\frac{1}{3}\log_2 \frac{1}{3}) + \frac{1}{4}(-\log_2 1) = \frac{3}{4}\log_2 3 \approx 1.189bits$ - asking this question is suboptimal as it provides[19] only $0.811bits$ of information about the state of the coins. Three quarters of the time Bob potentially needs another two questions to know exactly what the state of both coins are.

[18] We take the maximum entropy as the coins, and their tossing by Alice, are at least thought to be unbiased by Bob

[19] I use 'provides information' in the sense that the ignorance of Bob as measured by the entropy of his subjective probabilities is reduced.

The maximum information from a yes or no question is one bit - which happens when the answer evenly partitions the number of states (assuming each one is equally likely). The proof of this is not hard, but the result is intuitive enough for me to omit it. Of course, there are other questions that could be asked which would be optimal, such as "do both the coins show the same face?" Naturally, succeeding questions must chosen so as to reflect the first question (and sometimes its answer). In the example here, we see that Bob can gain complete knowledge about the state of the two coins by asking two of the right binary questions.

Given the possible states a system (here, the faces shown by some coins) and the nature of the questions (in the case above, yes or no questions), we can say how many questions are needed to exactly determine the state of the system.

However, there is a difficulty here. We are required to state what exactly the beliefs are about. This may be acceptable when an agreement between people can be made, but it is not when we talk of animals, or people, whose ontology is unknown to us.

An Thought Experiment Demonstrating a Problem with of Entropy

It is very clear that entropy is problematic in the case of continuous distributions, which may be easily demonstrated by a change of coordinates (see Shannon, 1948, Part IV). In the discrete case it is often claimed that problems concerning the ontology (or measures of it) have no bearing.

Carol, an excellent musician with perfect pitch, is the subject of an experiment about the confidence in which she can identify certain pieces of music. Carol is not formally trained and never concerns herself with classical music – indeed she could not name a single composer. The experimenter, Dave, tells her that he is going to play short excerpts from Johan Sebastian Bach's *Das wohltemperierte Klavier* and that it contains a one piece of music in each key. Dave plays a two second excerpt from one of Bach's preludes and fugues as establishes what probability Carol assigns to each[20]. He then calculates the entropy of her knowledge:

$$H_{\text{Carol}} = -\sum_k p_k \log p_k \ , \quad k \in \{Cmaj, C\#maj \ldots B\flat min, Bmin\} \tag{3.3.2}$$

Claire is just like Carol, in fact, she is a perfect copy of Carol up to the point where the experiment begins. Dave repeats the experiment, only afterwards, he tells Claire that *Das wohltemperierte Klavier* contains both preludes and fugues. Not knowing what a prelude

[20]Somehow. Methods exist for this, such as De Finetti's game, though in this example I shall just assume it is possible to do so whilst maintaining the procedure I describe.

or fugue is, the probability that it is a prelude or a fugue is one half. The probabilities for Claire are halved, but there are twice as many things to have the knowledge about. Dave now assigns an entropy to Claire's knowledge, as:

$$\begin{aligned} H_{\text{Claire}} &= \underbrace{-\sum_k \tfrac{1}{2}p_k \log \tfrac{1}{2}p_k}_{\text{Preludes}} \underbrace{-\sum_k \tfrac{1}{2}p_k \log \tfrac{1}{2}p_k}_{\text{Fugues}} \\ &\quad \text{so that } k \in \{Cmaj, C\#maj, Dmaj \ldots B\flat min, Bmin\} \end{aligned}$$

$$= -\sum_k p_k \log \tfrac{1}{2}p_k = \log 2 - \sum_k p_k \log p_k \tag{3.3.3}$$

$$= H_{\text{Carol}} + \log 2 \tag{3.3.4}$$

which is a difference of one bit of uncertainty (no knowledge about whether it is prelude or fugue). An interpretation of this in the spirit of Jaynes (1994) would be that it is in fact the entropy about Dave's beliefs, based upon Dave's ontology, when Dave does the calculation and that Dave, who is clear about what the ontology is used in each case, should understand the difference between the two. This is all well and good, but such an interpretation does not tell us about Carol and Claire's beliefs. Carol and Claire's beliefs are clearly the same in both cases, and really, we should expect to get an answer to that effect. The entropy here is sensitive to the ontology, really, what we require is a measure that is not sensitive in this way (measure invariant). No entropic measure that is of the form:

$$H^{(f)} = \mathbb{E}\left[f(x)\right] = \sum p(x) f(p(x)) \tag{3.3.5}$$

avoids this.[21] One way to avoid this problem is to make a measurement that uses our understanding of the ontology to cancel out its effects, or, in a more Jaynesian spirit, get Dave to report a quantity that takes this effect into account.

There are two ways to do this, using enumeratory measurements such as the channel capacity, or diversional measurements, such as the KL-divergence (Kullback and Leibler, 1951). I cover these in the next two section.

[21] Consider the transformation so that the ontology changes scale uniformly so that its size changes with $n \to m$ and the probabilities change under the maximum entropy rule such that $p_i \to q_i = \frac{n}{m}p_i$. Say that the entropy is changed as $H_n \to H_m$, then for invariance we require $H_n = H_m$ so we have $\sum_i p_i f(p_i) = \frac{m}{n}\sum_i q_i f(q_i) = \frac{m}{n}\sum_i \frac{n}{m}p_i f(\frac{n}{m}p_i) = \sum_i p_i f(\frac{n}{m}p_i)$ and therefore we need $f(p) = f(kp)$. The only solution to this is where f is a constant, meaning that the entropy it measures would also be constant and therefore rather pointless. Indeed, we cannot reasonably call a number that does not vary with anything a measurement.

3.3.3 Enumeratory Measures

Enumeratory measurements are, roughly, those that count something. Trivial examples may be counting the number of eggs in a basket, or the number of needles on a spruce tree. But also, lengths, volumes and areas can be thought of in this sense too. In essence, a given area can be thought of as being formed of multiple, infinitely small areas joined together. Notions such as length or area can be thought of as counting something that is infinite by giving each of those things an infinitesimal value.

Shannon's channel capacity (Shannon, 1948) can be thought of as an enumeratory measurement too. It, in a manner very much like counting the questions in the coin example, counts a number of distinguishable states that a communications channel can be in during a given time period. In this sense, it measures the number of elements an ontology must have to completely describe an set of given things and is geared to the in the case where there is noise that hinders doing so with the degree of accuracy of a deterministic scenario.

Let's say Carol was asked to communicate her best guess at the identity of the music to Dave, to do so well, she would need to use a channel with at least the capacity of $C_{\text{Carol}} = \log_2 24 = 4.58 bits$, whereas Claire, who is in addition required to specify whether it is a prelude or a fugue, would require the ability to transmit at least $C_{\text{Claire}} = \log_2(24 \times 2) = 5.58 bits$.[22] The channel capacity, like the entropy is dependant on the ontology, in fact, it is not just dependant on it, it *is* what it measures. By talking about channel capacity we are in effect postulating an object, the channel, the number of properties of which are either known or unknown. By measuring a channel capacity in a system, we are putting a lower bound on the number of properties this object has – the size of the ontology. If for example, we say the channel capacity is $10 bits$ we are in effect saying there is 1024 different states in our ontology.

3.3.4 Diversional Measures

Diversional measurements, in a sense, are the opposite of Shannon's channel capacity. They do not seek to measure the part of the entropy which is dependant on the ontology, but the part which isn't. They are concerned with information in the truest sense - how different are two beliefs. We can measure how much 'information' is required to change an

[22] The channel capacity can be written $\sup_{p(x)} MI(X, Y)$ and in the case of a discrete noiseless channel and a single message we can show that it is equal to $\log n$ where n is the number of different messages that we may wish to send.

agents belief from this to that, with minimal thought about how that belief is represented to that agent.[23] Formally, divergences compare two probability measures in a way that does not depend on the ontology. Using one of the principles that state that complete ignorance can be thought of as a probability distribution with equal probabilities[24] (in this case $\frac{1}{24}$) we can write a measure of the magnitude of difference between Carol's actual belief and complete ignorance as:

$$d_{\text{Carol}} = D_{\text{Carol}}\left(\left(\tfrac{1}{24}, \tfrac{1}{24} \cdots \tfrac{1}{24}\right) \;\middle|\middle|\; (p_{Cmaj}, p_{C\#maj} \cdots p_{Bbmin}, p_{Bmin}) \right) \tag{3.3.6}$$

A special case of which would be where D refers to the KL-divergence[25] I will define the KL-divergence properly later, but for now, I will just write it's value for the case above:

$$d_{\text{Carol}} = \sum_k p_k \log \frac{\left(\frac{1}{24}\right)}{p_k} = \log 24 - \sum_k p_k \log p_k \tag{3.3.7}$$

and for Clair we have

$$d_{\text{Clair}} = 2\sum_k \tfrac{1}{2} p_k \log \frac{\left(\frac{1}{48}\right)}{\frac{1}{2}p_k} = \log 24 - \sum_k p_k \log p_k \tag{3.3.8}$$

so

$$d_{\text{Carol}} = d_{\text{Clair}} \tag{3.3.9}$$

which is what we would sensibly expect from an information measure. Divergences measure only the 'knowledge part' of entropy, whereas channel capacities only measure the 'existence part' of the entropy. This comes at the cost only being able to compare two beliefs, where in entropy we have a way of describing a single belief. However, this is not so strange. The absolute position of points in space are really points relative to another point: the origin – there is no reason to expect it should be different in the case of diversional measures.

I should probably point out, that one can obtain a similar invariance by subtracting channel capacities from the entropies. In an informal sense, the entropies are measures which combine channel capacities with divergences.

[23] Though we do require that there is *sufficient* capacity to do so. This idea is formalised by Fisher Sufficiency.

[24] Such as the Principle of Maximum Entropy, or the Principle of Indifference *etc.*- they're all the same in *practice*.

[25] The KL-divergence has an interpretation in the spirit of Shannon, as how many bits of data would I need to transmit to change the receivers belief from the distribution on the left (right) to the distribution on the right (left) assuming I am using an encoding optimal for transmitting data according to the distribution on the left (right). But I wish to avoid using such interpretations in general.

3.4 The Logarithm in Information Theory

The logarithm is a reoccurring function in information theory, the examples above have used nothing else. I would like to take some time describing where they occur, why they occurred where they did and when and when not they are essential. It is good to make this explicit as it will allow me to both explain the motivation behind the approach used in later chapters and how information theory fits with what I have described so far.

Traditionally, logarithms occur throughout information theory, for example the Shannon entropy (Shannon, 1948) can be written as the expectation of the surprisal ($-\log p(x)$):

$$H(X) = \mathbb{E}\left[-\log p(x)\right] \tag{3.4.1}$$

The most widespread divergence, the KL-divergence (Kullback and Leibler, 1951) is also the expectation of a logarithm (see appendices A.1 and A.3 or chapter 4 for an explication of the notation):

$$D^{(KL)}\left(\xi \mid\mid \rho\right) = \mathbb{E}_{\xi}\left[\log \frac{p(x;\xi)}{p(x;\rho)}\right] \tag{3.4.2}$$

The Fisher score and Fisher information are defined as :

$$S_i = \frac{\partial}{\partial \xi^i} \log p(x;\xi) \tag{3.4.3}$$

and

$$g_{ij}^{\texttt{Fisher}} = \mathbb{E}_{\xi}\left[\frac{\partial}{\partial \xi^i} \log p(x;\xi) \frac{\partial}{\partial \xi^j} \log p(x;\xi)\right] \tag{3.4.4}$$

respectively, and the channel capacity also contains a logarithm (note the base):

$$C = \int_0^{\infty} \log_2\left(1 + \frac{S(f)}{N(f)}\right) \mathrm{d}f \tag{3.4.5}$$

In all these, the original motivation was because the quantities under consideration change geometrically. In the case of the channel capacity, entropy and the KL-divergence, it is because the number of different messages that one can represent with a sequence of symbols increases geometrically with the sequences length (Shannon, 1948). In the case of Fisher, it was motivated (at least in part) by the geometric increase in population sizes (Fisher, 1930).

Whilst the logarithm is traditional and has many nice properties, it is by no means necessary. For example, there are a number of alternatives for the entropy, such as those suggested by Rényi (1960) or Tsallis (2002). The class of divergences is also incredibly

large (Amari and Nagaoka, 2000), after all, all we really need is a way of comparing probability distributions with each other in a sensible way.

In contrast to the use of probability theory itself, where there is good reason to claim a correspondence to rational beliefs, there is no requirement to use logarithms (with one exception which I will discuss shortly). The bulk of chapter 4 explores other ways of measuring probabilities in a way so that they justified in the terms of rational goal-oriented behaviour.

In appendix C.1 we see that a logarithmic quantity arises naturally in the case of the Fisher metric. In many ways it is quite a natural choice. Indeed, the Fisher metric is the second derivative (curvature) of the Shannon entropy and the covariance matrix of the Fisher score. In terms of the justifications here, not their original justifications, they are natural because they ultimately come from local approximations of non-logarithmic divergences. With the philosophical grounding I have provided here, it is only the Fisher metric that has any real need for a logarithm.[26]

3.5 Summary: The Application of Information Measures

The focus of the first part has been on how we can justify using our own beliefs about the states of physiological components of an organism to discuss the *organisms beliefs*. The second part has been a review of various information theoretic measures. I will briefly review what I have said in the first part and show how it relates to the second.

From an information theoretic point of view, the biggest problem is the ontology. Firstly, our ontology should be sufficiently detailed to encapsulate the knowledge of the organism. As we do not know what the knowledge of the organism is about, we must make sure that our ontology contains the ability to represent all the situations which the organism's behaviour could be contingent on. This is the pragmatic maxim. A way of determining the size of ontology we need might be to measure channel capacity between some low level components. For example, if the channel capacity of the optic nerve is, say, $1Gbit/s$, then we know that we can represent the organism's belief concerning a given second of optical data with an ontology of any size greater than $2^{1,000,000,000}$ states.

We must also choose information measures that are prejudiced by the ontology. This, in addition to their natural occurrence in chapter 4, is where we find a need for divergences. As the size of the organisms ontology is not determined, we must choose something that is

[26] ...and even that is simply a consequence of choosing a geometry that is locally Euclidean! i.e. considering the covariance matrix of the fisher score, rather than some other statistic.

not affected by it. This comes at the cost of only being able to compare different beliefs, not giving them absolute numbers.

In addition, how we use our assessment of probabilities to describe the beliefs of other organisms is also dangerous. We must be very careful to ensure we are justified in doing so. We do this by arguing that our knowledge is equal to or greater than that of the organism so that the organism *cannot possibly* know anything more than we are aware of.

3.5.1 The Model Used in the Following Chapters

I have summarised the approach I will use in chapter 4 in figure 3.2. This figure should be approached cautiously as it is only a rough schematic, with labels chosen for brevity rather than exactitude.

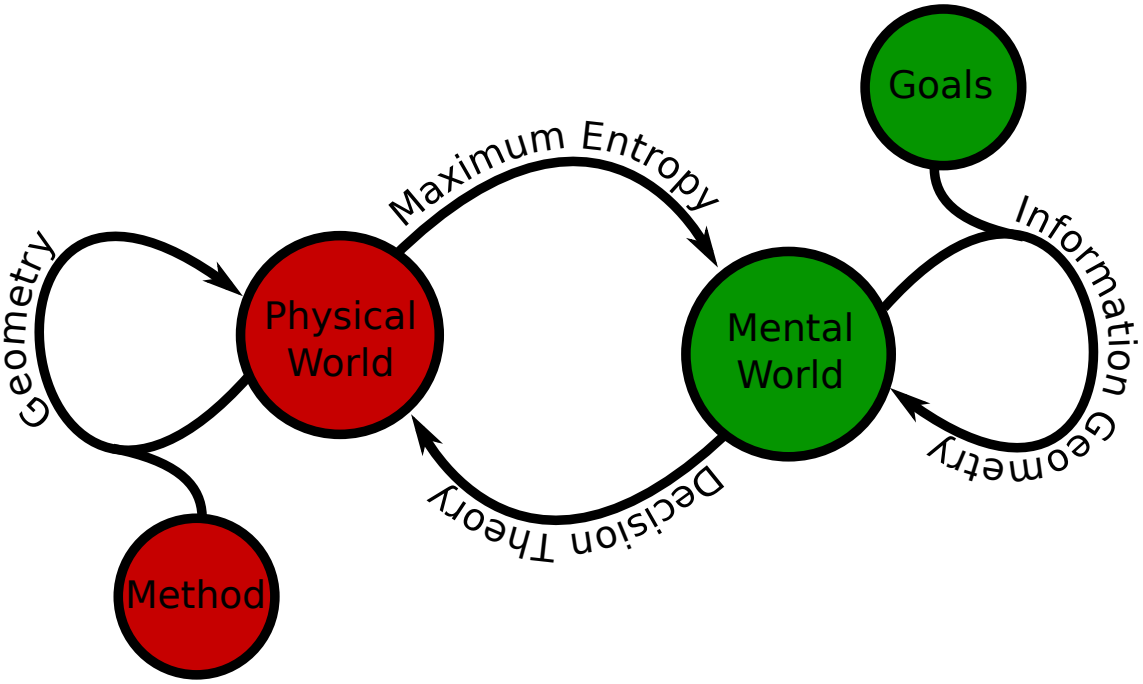

Figure 3.2: An informal diagram representing psychophysical relationships and quantification.

The diagram shows physical measurements on the left in red. These are geometric and measure the world as conceived of by modern physics. We use specific tools and methodologies to define geometric relationships between its components. On the right in green, we have beliefs (mental world), these can also be measured geometrically *someone with access to a goal of the organism* by judgements of the consequences in achieving that goal. We formalise these measurements using information geometry. The arrows between red and green are a possible way, but not the only way, we can relate the two domains. The labels used here are the formal tools that I will use frequently throughout the remaining

chapters.

The symmetry of this diagram is important. It shows the reason why I have made this case precisely. In physics we have method which allows us to understand exactly which measures to be making, but in information theory, there is presently not explicit rationale for choosing any one measure above another. I suggest, that it is the goals of the organism that provide us with exactly that. It is interesting to note that it could be said that the stance of some positivists is that the green side should just be a copy of the red side, and also, that it could also be said that it the stance of constructivist epistemology is that the red side should really just be another copy of the green side. In this sense, the diagram shows my stance as ratifying the two to some extent.[27]

To my knowledge there are no treatments of perceptual metrics that use this as a basis, though there is some similarity with signal detection theory. As we discover more and more about information theory we understand our motivations better: in the next chapter I will use the concepts above to explain the rationale behind colour metrics.

[27] Although, I would probably take the side of the constructivists if I had to make a stand.

Chapter 4

Geometry from Information

"The perception of what a thing is and the perception of what it means are not separate"

James J. Gibson, 1971
(Reed and Jones, Reasons for Realism)

This section outlines the origin of a Riemannian geometry as a description of colour. The derivation uses recent advances in theoretical machine learning and the relationship between information and geometry.

The derivation proceeds from a description of optimal classification - firstly providing a metric (the variational distance) which corresponds directly to the correctness of a classification. This distance measure is shown to correspond to a particular risk function - the 0-1 risk. From here I use results from machine learning as described by Bartlett et al. (2004); Nguyen et al. (2009) (and similar work in operations research by Jose et al. 2008) to describe a set of risk functions whose optimisation yields the same classification, and with them, the set of distance-like measures which correspond to their optimised value. This set of distance like measures (f-divergences) can all be thought of as sensible ways of judging long range colour differences. Indeed, it is by connecting the statistical theory with risk that we have a description of action, and thus, perceptually relevant quantities (O'Regan and Noë, 2001; Philipona and O'Regan, 2006; Rachlin, 1992).

From here, I use the information geometric techniques developed by Amari and Nagaoka (2000) and others to derive the appropriate Riemannian metric associated with these distance measures. Then in the next chapter I will show how this is equivalent to the spaces used in colour science.

Sketch of the Derivation

As I have suggested above, the following is a synthesis of three quite developed notions, behaviour, risk and information geometry. It is unlikely that the average reader will be well versed in all of these, so I have attempted to avoid an account that requires a lot of technical knowledge. For similar reasons, before I continue I will bullet point the derivation for reference:

1. Establish a standard behavioural experiment (2AFC) and standard error probabilities.
2. Show how this can be interpreted in terms of risk.
3. Show that a whole class of risk measurements give the same behaviour in the experiment.
4. Find corresponding information measures (divergences) - noting that they are symmetric.
5. Find their associated geometric quantities, showing that symmetry is preserved.
6. Observe that divergences do not necessarily lead to a Riemannian geometry.
7. Show that a standard Riemannian geometry is none the less generated from the risk based information measures as defined above (from symmetry).

The rest of the chapter will focus on the formal relationships to the general principles used in standard colour theory and extensions to more complex cases involving asymmetric risks and unequal prior probabilities.

Notation and Technicalities

Because this chapter is effectively the integration of two existing theories, there are two sets of notation which we will need to keep track of. The general set up of the probability distributions in the chapter is represented graphically in figure 4.1.

Throughout, I will use $x \in \mathcal{X}$ to describe some observed variable. I will take it that the distribution of the random variate X, on $\mathcal{X}$, is the marginal distribution of paired variables from another distribution on $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = \{-1, 1\}$.

I take these distributions to be parametrisable with parameters belonging to Ξ. I will then associate the *point* in parameter space $(\xi^1, \ldots, \xi^i)$ with the *class* where $y = 1$

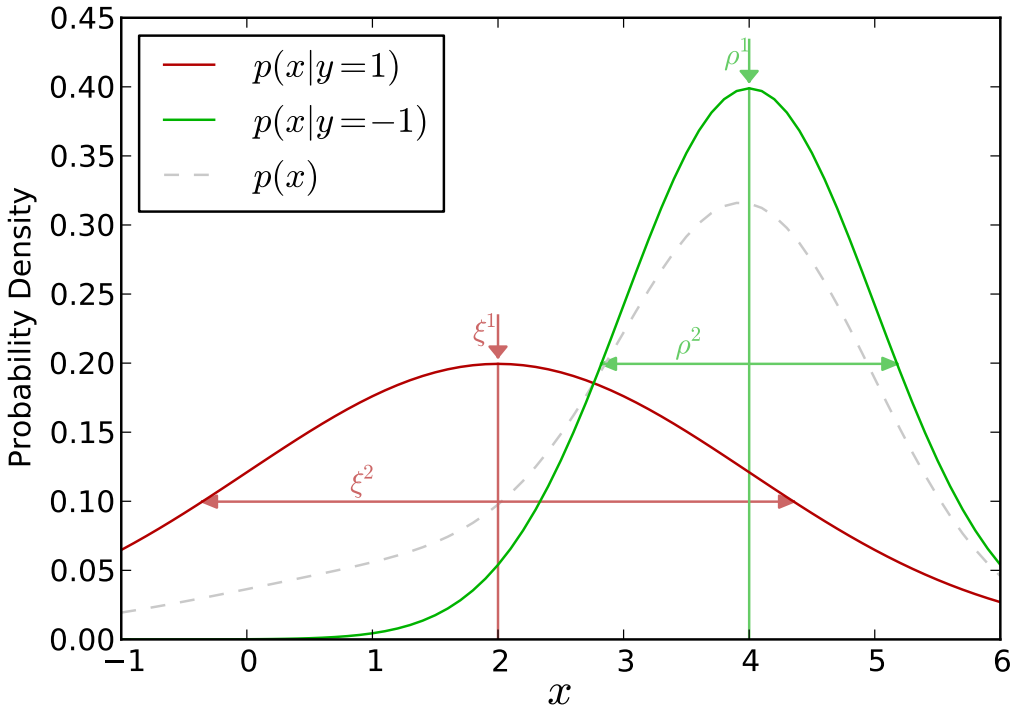

Figure 4.1: The structure of the probability spaces used in this chapter. The joint probability density $p(x, y)$ is composed of two densities, one for each $y \in \{-1, 1\}$. These are the two curves shown here. These two densities are parametrised by two variables $[\xi]$ and $[\rho]$, these are considered do define points within a statistical manifold.

and $(\rho^1, \ldots, \rho^i)$ with $y = -1$. I denote the probability densities of the parametrised distributions by $p(x; \xi)$ and $p(x; \rho)$. This means that the following holds:

$$\begin{aligned}
\Pr(x \in E \,|\, y = 1) &= \int_E p(x;\xi)\mathrm{d}\mu(x) \\
\Pr(x \in E \,|\, y = -1) &= \int_E p(x;\rho)\mathrm{d}\mu(x) \\
\Pr(y = 1) &= \tau_\xi \\
\Pr(y = -1) &= \tau_\rho \\
\Pr(x \in E,\ y = 1) &= \tau_\xi \int_E p(x;\xi)\mathrm{d}\mu(x) \\
\Pr(x \in E,\ y = -1) &= \tau_\rho \int_E p(x;\rho)\mathrm{d}\mu(x)
\end{aligned} \tag{4.0.1}$$

where μ is a measure continuous with the probability measure induced on $\mathcal{X}$ by conditioning the joint probability on $\mathcal{X} \times \mathcal{Y}$ upon members of $\mathcal{Y}$.[1]

[1] As long as we are careful with sets of measure zero this is perfectly fine. See Kullback and Leibler (1951) for more details.

There is also the matter of risk functions. In the case of loss and risk functions more generality can be achieved by extending their domain to the extended real line ($\mathbb{R} \cup \{\infty\}$). I will not need to invoke this myself at any point, but it is co-domain used in work on which I expand (specifically Nguyen et al., 2009).

4.1 2AFC and the Variational Distance

Before introducing the information geometric method upon which this chapter is based, I will first introduce the standard signal detection framework in a manner which is consistent with the rest of this chapter.

Consider a subject of an experiment who is forced to make a choice between two alternatives (a and b). The two alternatives have properties ($x|a$ and $x|b$, the physiological state given an alternative) which are drawn from two distributions ($p(x|a)$ and $p(x|b)$). Let's assume that the subject's motivation and ability to correctly assign x to either a or b is ample. Then the subjects best strategy in terms of minimising the probability of error is to assign a whenever the inequality $p(x, a) > p(x, b)$ holds. Here we take the maximum entropy distribution for the prior distribution of alternatives a and b (Jaynes, 1994):

$$p(a) = \arg\max_p \{p \log p + (1-p)\log(1-p)\} \tag{4.1.1}$$

$$= \tfrac{1}{2} = 1 - p(a) = p(b) \tag{4.1.2}$$

so

$$p(x, a) = \tfrac{1}{2} p(x|a) \text{ and } p(x, b) = \tfrac{1}{2} p(x|b) \tag{4.1.3}$$

Given this, we can then write an equation for the expected error:

$$\epsilon = \int_{\mathcal{X}} f(x) \mathrm{d}\mu(x) \tag{4.1.4}$$

where

$$f(x) = \begin{cases} p(x, b), & p(x, a) > p(x, b) \\ p(x, a), & \textit{otherwise} \end{cases} \tag{4.1.5}$$

which can also be written:

$$f(x) = \tfrac{1}{2} \left(p(x, a) + p(x, b) - |p(x, a) - p(x, b)| \right) \tag{4.1.6}$$

and thus

$$\begin{aligned}\epsilon &= \tfrac{1}{2}\int_{\mathcal{X}} p(x,a)+p(x,b)\mathrm{d}\mu(x) - \tfrac{1}{2}\int_{\mathcal{X}} |p(x,a)-p(x,b)|\,\mathrm{d}\mu(x) \\ &= \tfrac{1}{2} - \tfrac{1}{2}\int_{\mathcal{X}} |p(x,a)-p(x,b)|\,\mathrm{d}\mu(x) \\ &= \tfrac{1}{2} - \tfrac{1}{4}\int_{\mathcal{X}} |p(x|a)-p(x|b)|\,\mathrm{d}\mu(x) \\ &= \tfrac{1}{2} - \tfrac{1}{4}V(p(x|a),p(x|b)) \end{aligned} \tag{4.1.7}$$

and the probability of success, s, is simply equal to:

$$s = 1-\epsilon = \tfrac{1}{2} + \tfrac{1}{4}V(p(x|a),p(x|b)) \tag{4.1.8}$$

$V(\cdot,\cdot)$ is known as the (total) variational distance. It should be noted that the quantity:

$$\int_{\mathcal{X}} |p(x,a)-p(x,b)|\,\mathrm{d}\mu(x) \tag{4.1.9}$$

is also known by this name, being more general in the sense that it accounts for any prior probabilities $p(a)$ and $p(b)$. A generalisation for asymmetric error weightings as well as prior probabilities is given in appendix B, this generalisation is of some relevance to the extension in section 4.7.1.

The variational distance plays an important role in information theory and its relationship with informational quantities such as the Kullback-Leibler Divergences are numerous (e.g. Lin, 1991; Topsœ, 2000). The variational distance takes a value in $[0,2]$ and is a metric on the space of probability distributions. It is immediately obvious from this that the success takes a value in $[\frac{1}{2},1]$ as one would hope.

This distance corresponds very well to the standard psychometric curve, for good reason - the standard psychometric curve can be thought of as the variational distance (or an approximation to it) under particular assumptions. The sigmoid shape is a general feature of signal detection problems.

Expression in Terms of Bayes Risk

The formulation above can be thought of in terms of the binary case of Bayes risk. Bayes risk is related to the variational distance - it is almost the same. However, how one arrives at the Bayes risk is slightly different. Importantly, the usual set up for Bayes risk includes a functions known as the discriminant function, γ, and the loss function ϕ. The derivation, in the spirit of (Nguyen et al., 2009) is as follows:

Similar to above, we begin with a distribution $(\mathcal{X},\mathcal{Y})$ of observed properties $\mathcal{X}$ and corresponding classes $\mathcal{Y}$, one can think the pair $(x,y)\in(\mathcal{X},\mathcal{Y})$ as a pair of observed values

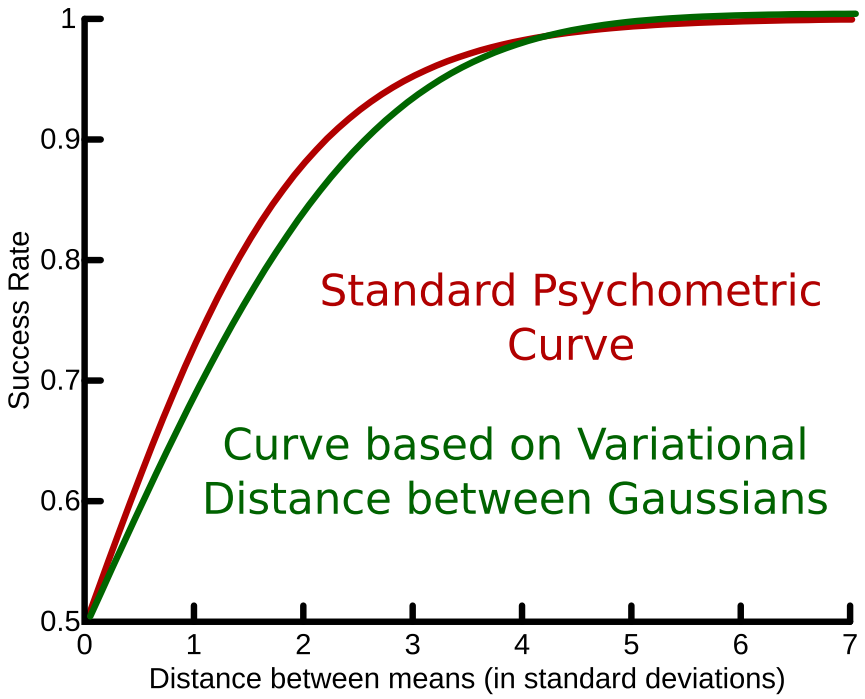

Figure 4.2: The "success" for Gaussian variables as used in ROC analysis, compared to a psychometric curve ($1/(1+\exp(-x))$) which would be obtained using logistic regression.

(x) with the correct class to assign it (y). To make things easier, we assume that the labels of $y \in \mathcal{Y}$ are either -1 and 1 (instead of a and b above). Next, we define a discriminant function γ which assigns a real number to each possible $x \in \mathcal{X}$. The sign of this function is used to predict the correct value of y from x, so that we have an estimation of y, called $\hat{y}$:

$$\hat{y}(x) = \operatorname{sgn} \gamma(x) \tag{4.1.10}$$

so that $\hat{y}$ almost surely[2] belongs to $\{-1, 1\}$. As both y and $\hat{y}(x)$ take values of either -1 or 1 it is obvious that:

$$y\hat{y}(x) = \begin{cases} 1, & y = \hat{y}(x) \\ -1, & otherwise \end{cases} \tag{4.1.11}$$

and that (almost surely):

$$y = \hat{y}(x) \implies y\gamma(x) > 0 \tag{4.1.12}$$

$$y \neq \hat{y}(x) \implies y\gamma(x) < 0 \tag{4.1.13}$$

[2]The difficulty of $\operatorname{sgn} 0 = 0$ is often remedied by letting $\operatorname{sgn} 0 = 1$ or $\operatorname{sgn} 0 = -1$. However, this problem can usually be ignored.

From, here it is easy to write a function that gives us the probability of an error, which I shall call the Bayes risk.[3] Letting $\mathbb{I}(\varphi)$ be the indicator function for a proposition φ:

$$\mathbb{I}(\varphi) = \begin{cases} 1, & \varphi \text{ is true} \\ 0, & \textit{otherwise} \end{cases} \tag{4.1.14}$$

This is used to define the quantity known as 0-1 loss ($\phi_{0\text{-}1}$). It's expectation is the Bayes risk.

$$\phi_{0\text{-}1}(\beta) = \mathbb{I}(\beta < 0) \tag{4.1.15}$$

then we have:

$$\begin{aligned} \mathcal{R}_{\text{Bayes}}(\gamma) &= \mathbb{E}\left[\phi_{0\text{-}1}(y\gamma(x))\right] && (4.1.16)\\ &= \mathbb{E}\left[\mathbb{I}(y\gamma(x) < 0)\right] && (4.1.17)\\ &= \int_{(\mathcal{X},\mathcal{Y})} \mathbb{I}(y\gamma(x) < 0)\, p(x,y) \mathrm{d}\mu(x,y) && (4.1.18) \end{aligned}$$

in words, this is the probability of making a mistake - the chance that γ has the wrong sign.

We are now positioned to relate the Bayes risk to the variational distance. It should be clear that with the probabilities of y values being equal, we get the following relationship with the variational distance:

$$\inf_{\gamma} \mathcal{R}_{\text{Bayes}}(\gamma) = \epsilon = \tfrac{1}{2} - \tfrac{1}{4} V(p(x|a), p(x|b)) \tag{4.1.19}$$

This relationship will be generalised in the next section to eventually yield information geometric quantities.

4.2 Bayes Consistent Risk Functions

In the expression of risk I have discussed, there is the scope to describe the risk associated with uncertainty. The value in reducing uncertainty is the major determinent of the results here. This is known in economics as uncertainty aversion. Standard information theoretic quantities can be thought of as describing uncertainty aversion where the risk is linear with message length or some similar quantity.

Since we are only concerned with the sign of γ there is an uncountable infinity of definitions of $\gamma(x)$ which yield the same classification $\hat{y}(x)$, i.e. the mapping $\gamma \to \operatorname{sgn}\gamma$ is many to one.

[3] Though often Bayes risk is considered to be the general formulation of risk involving the expectation of a loss function.

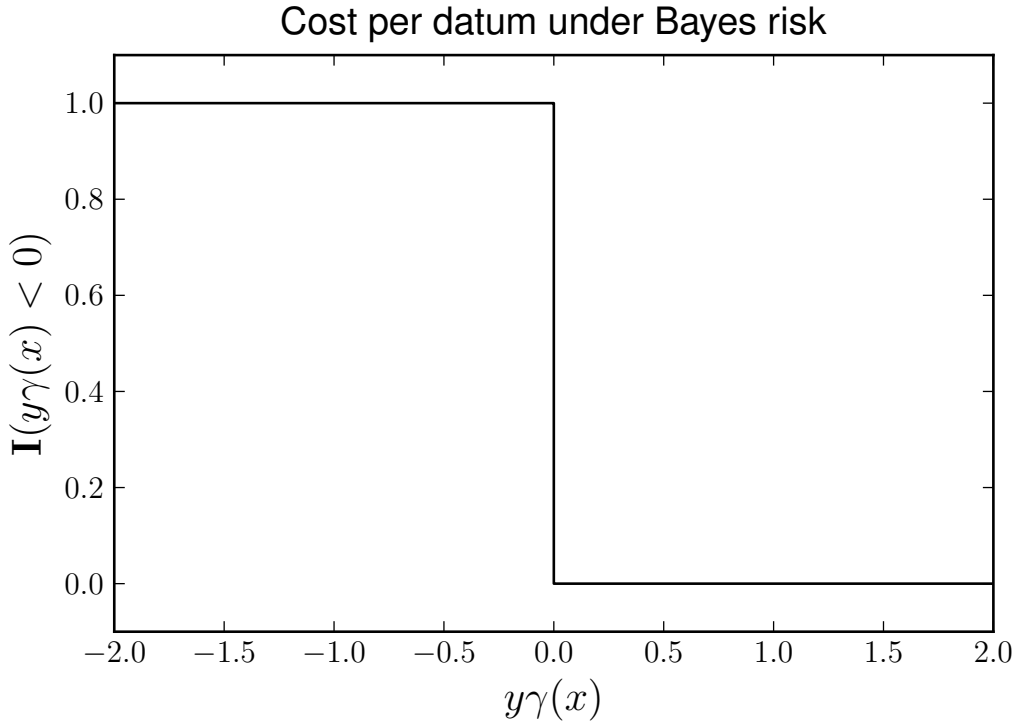

Figure 4.3: Plot of the 0-1 loss ($\phi_{0\text{-}1}$) with respect to the output of the discriminant function. There are two things to note: Firstly, the magnitude of the discriminant function does not influence the result, only it's sign. Secondly, the function is not convex.

This section focuses on risk other than the Bayes risk, with the ultimate aim of describing risk measures which yield the same $\hat{y}$. In other words, what other ways are there defining risk which yield the same classification as the Bayes risk. This question has been answered by Bartlett et al. (2004) in the general case and by Lugosi and Vayatis (2004) for the specific case of boosting.

The Bayes risk used the function $\mathbb{I}(\alpha < 0)$ (where $\alpha = \gamma(x)$) to map the discriminant function to a contribution to risk, but there are numerous (infinite) ways of doing so, represented in general by ϕ. I shall focus on a collection of functions often called surrogate loss functions, so named because they are used in machine learning as tractable surrogates for the 0-1 loss of the Bayes risk. These are not the complete repertoire of possible loss functions, but they constitute a large class of well behaved functions that is sufficient to demonstrate the the major points of this section. Surrogate loss functions are convex upper bounds on the 0-1 loss - they, by definition, must be convex[4] and satisfy $\phi(w) > 1$ when $w \leq 0$ and $\phi(w) \geq 0$ when $w > 0$. Whilst any loss function that satisfies this could

[4] Allowing us to use convex analysis. This is what makes them tractable surrogates!

rightly be called a surrogate loss function, one would be missing a trick not replacing the two inequalities with the requirement that $\phi(0) = 1$ (and $\phi(x) \geq 0$). This assures that ϕ bounds $\phi_{0\text{-}1}$ as tightly as possible. First of all, these functions are functions only of $y\gamma(x)$, a class of functions referred to as marginal loss functions ($y\gamma(x)$ is known as the margin). The ϕ-risk is then defined as:

$$\mathcal{R}_\phi(\gamma) = \mathbb{E}[\phi(y\gamma(x))] \tag{4.2.1}$$

Also, we define the optimal ϕ-risk. This is the minimal risk that can be obtained given free choice of the discriminant γ.

$$\mathcal{R}_\phi^* = \inf_\gamma \mathcal{R}_\phi(\gamma) \tag{4.2.2}$$

A theorem of (Nguyen et al., 2009) is that, with some conditions on the nature of ϕ:

$$\mathcal{R}_\phi(\gamma) \to \mathcal{R}_\phi^* \implies \mathcal{R}_{\text{Bayes}}(\gamma) \to \mathcal{R}_{\text{Bayes}}^* \tag{4.2.3}$$

As the ϕ-risk associated with a discriminant tends to its minimum value, so does its Bayes risk. Or, more to the point, a ϕ-optimal discriminant function is Bayes optimal. The conditions alluded to above are simple for the convex ϕ losses, and can be summarised as:

$$\phi'(0) < 0 \tag{4.2.4}$$

which has the simple interpretation when taken along with the convexity: any possible error is *always* punished more than any correct choice.

4.2.1 Risk and Distance

Now that we know that the optimal ϕ-risk corresponds to a γ that is Bayes optimal we can move on to asking how this relates to what might be considered a colour-distance.

The following is a summary of the results of Nguyen et al. (2009) in the slightly narrower context of equal prior class probabilities, i.e. $p(y = -1) = p(y = 1) = \frac{1}{2}$. This makes it easy to speak of conditional probabilities. Relaxation of this constraint is straight forwards once we have established the information geometric formulation and will be described in section 4.7.2. For now, I shall proceed with the maximum entropy prior.

Here I will make use of the following notation, this is to aid the transition to the descriptions used in information geometry. Letting y_1 and y_2 represent the cases where $y = -1$ and $y = 1$ respectively:

$$p(x, y_1) = p(y_1)p(x|y_1) = \tfrac{1}{2}p(x;\xi) \tag{4.2.5}$$

$$p(x, y_2) = p(y_2)p(x|y_2) = \tfrac{1}{2}p(x;\rho) \tag{4.2.6}$$

This change of notation corresponds to adopting a new way of looking at the probabilities. Conceptually, we consider the probabilities $p(x;\xi)$ to belong to a statistical manifold. A statistical manifold is a space of probability distributions, parametrised by some set of parameters (coordinates) $(\xi^1 \dots \xi^N) \in \Xi$. Each point on the manifold, $p(x;\xi)$, is a distribution with support $\chi = \{x\}$ which is parametrised by ξ. Example in footnote[5].

The equation for ϕ-risk can be written in terms of this new notation.

$$\begin{aligned} \mathcal{R}_\phi(\gamma) &= \mathbb{E}\left[\phi\left(y\gamma(x)\right)\right] && (4.2.7)\\ &= \tfrac{1}{2}\int_\chi \left(\phi\left(\gamma(x)\right)p(x;\xi) + \phi\left(\gamma(x)\right)p(x;\rho)\right)\mathrm{d}\mu(x) && (4.2.8)\end{aligned}$$

Letting us minimise in terms of $\gamma(x)$. To do this we must note that minimising each of the individual elements of the integral minimises the integral itself. Letting $\gamma(x) = \beta$ for clarity:

$$\begin{aligned} \mathcal{R}^*_\phi &= \tfrac{1}{2}\int_\chi \inf_\beta \left[\phi(\beta)p(x;\xi) + \phi(-\beta)p(x;\rho)\right]\mathrm{d}\mu(x) && (4.2.9)\\ &= \tfrac{1}{2}\int_\chi p(x;\xi)\inf_\beta \left[\phi(\beta) + \phi(-\beta)\frac{p(x;\rho)}{p(x;\xi)}\right]\mathrm{d}\mu(x) && (4.2.10)\end{aligned}$$

Letting $u = \frac{p(x;\rho)}{p(x;\xi)}$ and:

$$f(u) = -\inf_\beta \left[\phi(\beta) + \phi(-\beta)u\right] \tag{4.2.11}$$

the optimal risk can be written as:

$$\mathcal{R}^*_\phi = \tfrac{1}{2}\int_\chi p(x;\xi) f\left(\frac{p(x;\rho)}{p(x;\xi)}\right)\mathrm{d}\mu(x) \tag{4.2.12}$$

The class of quantities described by this equation are known as f divergences.[6] The class of f-divergences contains all quantities that corresponds to:

$$\int_\chi p(x;\xi) f\left(\frac{p(x;\rho)}{p(x;\xi)}\right)\mathrm{d}\mu(x) \tag{4.2.13}$$

These are often written as $D^{(f)}\left(\xi \,||\, \rho\right)$ (Amari and Nagaoka, 2000), in this notation we have:

$$\mathcal{R}^*_\phi = -\tfrac{1}{2}D^{(f)}\left(\xi \,||\, \rho\right) \tag{4.2.14}$$

[5] For example, if we have a one dimensional manifold of Poisson distributions parametrised by their rate, then we have a manifold (Ξ, χ, P) such that $P = p(x;\xi) = \frac{\xi^x e^{-\xi}}{x!}$, $\Xi = \mathbb{R}_+, \chi = \mathbb{N}$

[6] There is usually a convexity requirement, but I will get to this later.

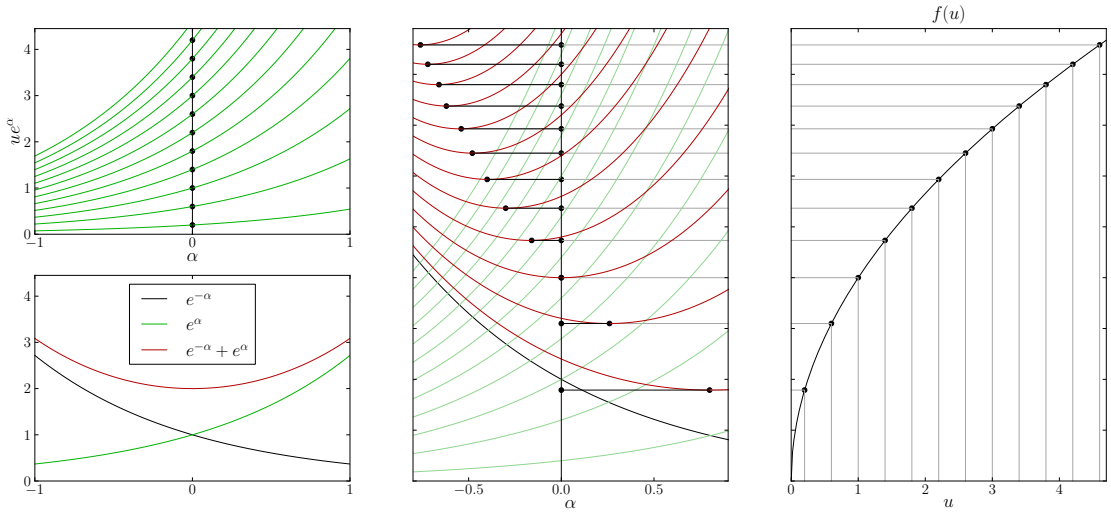

Figure 4.4: Example of the relationship between the ϕ loss and the convex function f of f-divergences. In this case I show the relationship in the case of the exponential loss: $\phi(\beta) = e^{-\beta}$. Beginning on the top left - the functions $u\phi(-\beta)$ for various u, the value of u in this case is found where $\beta = 0$. On the bottom left - the three functions $\phi(\beta)$, $\phi(-\beta)$ and $\phi(\beta)+u\phi(-\beta)$ for $u = 1$ (i.e. $\phi(\beta)+\phi(-\beta)$). The centre panel shows $\phi(\beta)+u\phi(-\beta)$ for evenly spaced u ($u = \{0.2, 0.6, 1, 1.2...\}$). Here I mark the minima (given by $\inf\{\phi(\beta) + u\phi(-\beta)\}$), the height of the minimum is the value of the divergence of a given u. The grey lines at each minimum continue on into the right panel, where the function $f(u)$ relates u to those heights. In this case $f(u) = 2\sqrt{u}$.

4.3 f-Divergences and h-Divergences

In the f-divergences, we have a broad class of functions that could be sensibly thought of as general measures of difference. All of these correspond to how well two stimuli can be distinguished. There are of course other functions, but the fact that there is more than one is of great significance. All of them correspond to making the same classification, so from the point of view of making the best possible choice they are the same. No one of them, when considering the classification to which it corresponds, is better than any other. All of them measure difference is a sensible way - how well can the best classifier do, given some way of quantifying of how well it is doing.

I should be noted that the general class of f-divergences is actually broader than those that correspond to a ϕ-risk as defined above. The f-divergences corresponding to a ϕ-risk are symmetric (Bartlett et al., 2004):

$$D^{(f)}(\xi \,||\, \rho) = D^{(f)}(\rho \,||\, \xi) \tag{4.3.1}$$

which does not hold for f-divergences in general. This will not be a problem as much of what follows will apply in general to any f-divergence, symmetric or not (I will be explicit when symmetry is used). However, I will impose a different constraint to these divergences. Namely, that the divergences considered will be zero for two points that are the same. Often this does not hold for divergences derived from ϕ-risks, but can simply be corrected by the addition of a quantity. This is in a sense a *geometrising* transform: it changes an f-divergence that does not obey certain axioms of metrics to one that does. Doing this makes it easier to apply the formalisms of Amari and Nagaoka (2000). Letting h correspond to the convex function defining the corrected divergence, we have:

$$h(u) = f(u) + (1-u)f'(1) - f(1) \tag{4.3.2}$$

in which h can easily be shown to be a convex function. The reason for the constant term is twofold: From the psychometric perspective, choosing a measurement where the "distance" between a point and itself is zero corresponds to our intuitive notion of distance. If we were to compare distances, asking is one bigger than another a additive constant would have little impact. Adding a constant does not affect the fact that larger f-divergences mean a better classification, moreover, it grounds them by stating that the difference between stimuli indistinguishable to a sufficiently able experimenter will always measure as zero. The other reason - from a mathematical perspective - is that not having zero distance from a point to itself does not produce anything that resembles what one would like to call a geometry. The term first order in u assures that the divergence is always positive.

Importantly, *this operation preserves the symmetry of the divergence* - this is demonstrated in the general case in appendix C.2. Another important property of this is that it does not affect the connection coefficient corresponding to the geodesics that connect two points (which I shall discuss soon). The geometrising transform adds a *constant* that is dependant on f and the prior probability of the two classes at each point, most generally we can write:

$$D^{(h)}\left(\xi \,||\, \rho\right) = D^{(f)}\left(\xi \,||\, \rho\right) + (\pi_\xi - \pi_\rho)f'(1) - \pi_\xi f(1) \tag{4.3.3}$$

where $\pi_\xi = p(y_1)$ and $\pi_\rho = p(y_2)$.

In the current case with equal prior probabilities (and where divergences are in a sense doubled) the geometrising transform is simply:

$$D^{(h)}\left(\xi \,||\, \rho\right) = D^{(f)}\left(\xi \,||\, \rho\right) - f(1) \tag{4.3.4}$$

More formally, the correction above assures the following axioms of metrics: the iden-

tity of indiscernibles

$$D^{(h)}\left(\, p \,||\, p \,\right) = 0 \leftrightarrow h(1) = 0 \tag{4.3.5}$$

and non-negativity

$$D^{(h)}\left(\, p \,||\, q \,\right) \geq 0 \leftrightarrow h'(1) = 0 \tag{4.3.6}$$

the exact form of this term can be derived from Taylor's expansion of the divergence (see appendix C.1). Symmetry of the divergence can also be written as a property of the divergence defining function, h:

$$D^{(h)}\left(\, p \,||\, q \,\right) = D^{(h)}\left(\, q \,||\, p \,\right) \leftrightarrow h(u) = uh(\tfrac{1}{u}) \tag{4.3.7}$$

Specific Cases

The variational distance is itself an h-divergence with $f(u) = |1 - u|$. However, importantly, it is not differentiable at $u = 1$. It is not differentiable exactly where we will need it to be when we want to differential geometry. Table 4.1 shows some common f-divergences, their common names and common corresponding ϕ-risks.

Divergence Class	$f(u)$	**Loss Name**	$\phi(\alpha)$
Variational Distance	$\lvert 1 - u \rvert$	0-1 Loss	$\mathbb{I}\left(\alpha < 0\right)$
		Hinge Loss	$\max\{0, 1 - \alpha\}$
Hellinger Distance	$(1 - \sqrt{u})^2$	Exponential Loss	$e^{-\alpha}$
Fisher-Rao Distance			
Triangular Distance	$\frac{-4u}{u+1}$	Least Squares Loss	$(1 - \alpha)^2$

Table 4.1: Some definitions of common f-divergences with some common ϕ-risks

4.4 Differential Manifolds

This chapter requires a basic understanding of differential geometry. The geometry used here is not the standard Riemannian (or pseudo-Riemannian) geometry that will be familiar to physicists, but a minor generalisation of it.

The most important object in differential geometry is the manifold. A manifold, M, is a continuum of points, here they are determined by some coordinates $\xi = (\xi^1 \dots \xi^N)$.[7]

[7]Not that any coordinate system is required at all.

The power of differential geometry lies in the fact that the manifold M is assumed to be the same no matter what coordinates are used, say $\rho = (\rho^1 \dots \rho^N)$ - it is coordinate invariant.

But what do I mean by "the same". The thing that needs to be the same is a particular notion of distance between points, not one as measured by some difference in coordinates, as this will change if we choose different coordinates, but by some other measure of distance, defined in addition to the coordinates - distances *intrinsic* to the manifold. This function (F) is a function of two points, but importantly only those pairs of points that are "next to" each other on the manifold. Making some assumptions about the continuity of the coordinates and the manifold, we can say that a point at ξ is "next to" a point at $\xi + \mathrm{d}\xi$, where $\mathrm{d}\xi$ is some infinitesimally small displacement in the coordinates. This is true for any coordinate system.

Then at each point on the manifold we define a the infinitesimal distance between neighbouring points, $\mathrm{d}s$, to be given by:

$$\mathrm{d}s = F\left(\xi^1, \dots, \xi^N, \mathrm{d}\xi^1, \dots, \mathrm{d}\xi^N\right) \tag{4.4.1}$$

Is is this distance, $\mathrm{d}s$, which is the intrinsic property that is kept the same. There are numerous sensible ways of defining F. Generally, as long as $F(\xi, k\mathrm{d}\xi) = kF(\xi, \mathrm{d}\xi)$ one gets something sensible. There is, however, a particular case which is widely used - popularised by Bernhard Riemann:

$$\mathrm{d}s = F(\xi, \mathrm{d}\xi) = \sqrt{\sum_i \sum_j g_{ij}(\xi)\mathrm{d}\xi^i \mathrm{d}\xi^j} = \sqrt{g_{ij}(\xi)\mathrm{d}\xi^i \mathrm{d}\xi^j} \tag{4.4.2}$$

The last equality uses notation that will be used throughout: when there are quantities with upper and lower indices next to each other summation is assumed (see appendix A.3 for a summary of the notation used throughout). The quantity $g_{ij}(\xi)$, usually just written g_{ij} is a function of the coordinate ξ. Objects like g defined over coordinates are known as *tensor fields* or just *tensors*[8]. In the case of g - which basically defines $\mathrm{d}s$ - it is known as the *metric tensor*.

4.4.1 Transformation of Coordinates

To maintain the same geometry in coordinates ξ and coordinates ρ is the same as requiring $\mathrm{d}s$ to be the same in both coordinates. This means that the metric tensor has a different

[8] usually, objects are only called tensors if they behave in a particular way under change of coordinates, see equations 4.4.3 and 4.4.4

representation in the two coordinate systems. The transformation between ξ and ρ should obey:

$$g_{ij}(\xi)\mathrm{d}\xi^i\mathrm{d}\xi^j = \mathrm{d}s^2 = g_{ab}(\rho)\mathrm{d}\rho^a\mathrm{d}\rho^b \tag{4.4.3}$$

and noting that $\mathrm{d}\xi^i = \frac{\partial\xi^i}{\partial\rho^a}\mathrm{d}\rho^a$ then

$$g_{ij}(\xi)\frac{\partial\xi^i}{\partial\rho^a}\frac{\partial\xi^j}{\partial\rho^b} = g_{ab}(\rho) \tag{4.4.4}$$

4.4.2 Geodesics and Long Distances

The distance between points, which do not neighbour each other requires integrating along a curve. A curve between points a and b has coordinates $\gamma(t) = (\gamma^1(t)...\gamma^N(t))$ and it is usually parametrised by a variable t such that $\gamma(0) = a$ and $\gamma(1) = b$. The length of the curve is then:

$$s(\gamma) = \int_a^b \mathrm{d}s = \int_{t=0}^{t=1} \sqrt{g_{ij}(\gamma(t))\left(\frac{\partial}{\partial t}\gamma^i(t)\right)\left(\frac{\partial}{\partial t}\gamma^j(t)\right)}\ \mathrm{d}t \tag{4.4.5}$$

Of course there is more than one curve between two points. Often we want the shortest curve, which is uniquely defined in Riemannian geometry as well as the extension here - we want to calculate $\inf_\gamma s(\gamma)$ for all gammas with $\gamma(0) = a, \gamma(1) = b$. These shortest curves are geodesics and are usually given by[9](Postnikov, 1998):

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\xi^i(t) + \Gamma^i_{jk}(\xi(t))\frac{\mathrm{d}\xi^j(t)}{\mathrm{d}t}\frac{\mathrm{d}\xi^k(t)}{\mathrm{d}t} = 0 \tag{4.4.6}$$

where Γ^i_{jk} are the (affine) connection coefficients.

Up to this point the geometry outlined is bog-standard. But with the geodesic equation we come across the assumption which will be relaxed. The most intuitive way of looking at how the geometry used here is to look at how we arrive at the geodesics. Usually, the calculation of the geodesics uses an energy functional of the form:

$$E(\gamma) = \tfrac{1}{2}\int_{t=0}^{t=1} g_{ij}(\gamma(t))\frac{\partial\gamma^i(t)}{\partial t}\frac{\partial\gamma^j(t)}{\partial t}\mathrm{d}t \tag{4.4.7}$$

whose minimisation ensures minimisation of s as

$$s(\gamma)^2 \leq 2E(\gamma) \tag{4.4.8}$$

which is obtained quite directly from the Cauchy-Schwartz inequality.

Now, when E is some other quantity, like one of the many derived from risk functions, these rules change – the arc-length may no longer minimise E. Yet, as we will see, the local geometry remains the same.

[9]This only applies locally. It is the shortest in the sense that there no shorter curves formed by changing it slightly. There may shorter curves if a big change to the curve is made.

4.4.3 Other Measures of Distance

In this chapter we are interested in other distance like measures than s. For the case of E and s there is a particular choice of Γ_{ij}^k for which equation 4.4.6 yields the correct curve. The values of Γ_{ij}^k are known as the Levi-Civita, or Riemannian, connection coefficients and are given by:

$$\Gamma_{ij,k} = g_{lk}\Gamma_{ij}^l = \tfrac{1}{2}\left(\frac{\partial}{\partial\xi^i}g_{kj} + \frac{\partial}{\partial\xi^j}g_{ik} - \frac{\partial}{\partial\xi^k}g_{ij}\right) \tag{4.4.9}$$

When not using E or s the connection coefficients Γ_{ij}^k are not necessarily these - indeed, there is a different connection for every real number (denoted α). I will explain what they are in the following sections.

Some of these choices yield f-divergences exactly, such as $\alpha = \pm 1$, (Amari and Nagaoka, 2000) in this case it is the Kullback leibler divergence for which we can roughly consider the risk to be linear with a reaction time (this follows from arguments similar to Shannon, 1948).

4.5 h-divergence to Fisher Information

With a class of distance measures defined, we can now look at the differential geometry that corresponds to them. To do this, we begin with the Taylor's expansion of h-divergences to the third order (this of course assumes the appropriate differentiability at $u = 1$):

$$\begin{aligned} D^{(h)}(\,\xi \,||\, \xi + \Delta\xi\,) \;=\; & D^{(h)}(\,\xi \,||\, \rho\,)\Big|_{\rho=\xi} + \\ & \frac{\partial}{\partial\rho^i}D^{(h)}(\,\xi \,||\, \rho\,)\Big|_{\rho=\xi}\Delta\xi^i + \\ & \frac{1}{2}\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}D^{(h)}(\,\xi \,||\, \rho\,)\Big|_{\rho=\xi}\Delta\xi^i\Delta\xi^j + \\ & \frac{1}{6}\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}\frac{\partial}{\partial\rho^k}D^{(h)}(\,\xi \,||\, \rho\,)\Big|_{\rho=\xi}\Delta\xi^i\Delta\xi^j\Delta\xi^k + \\ & o(\Delta\xi^4) \end{aligned} \tag{4.5.1}$$

The expansion is worked through in appendix C.1. Firstly, we consider the expansion up to and including $\Delta\xi^2$. This is simply:

$$D^{(h)}(\,\xi \,||\, \xi + \Delta\xi\,) = f''(1)\int_\chi \frac{1}{p_\xi}\frac{\partial}{\partial\xi^i}p_\xi\frac{\partial}{\partial\xi^i}p_\xi \mathrm{d}\mu(x)\Delta\xi^i\Delta\xi^j + o(\Delta\xi^3) \tag{4.5.2}$$

where $p_\xi = p(x;\xi)$. The integral in this can be written as:

$$\int_\chi \left(\frac{\partial}{\partial\xi^i}\ell_\xi\right)\left(\frac{\partial}{\partial\xi^j}\ell_\xi\right)p_\xi \mathrm{d}\mu(x) \tag{4.5.3}$$

which is usually written as an expectation[10]. Here $\partial_i = \frac{\partial}{\partial \xi^i}$ and $\ell_\xi = \log p(x;\xi)$ - a notation which will be used throughout this chapter:

$$g_{ij} = \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi\right] \tag{4.5.4}$$

This is known as the Fisher metric, or, Fisher information. At small enough $\Delta\xi$ we see that any divergence is proportional to the Fisher metric based arc-length squared.

4.5.1 The Fisher Metric

The Fisher metric (or information) is symbolised as $g_{ij}^{\texttt{Fisher}}$ - or when its identity is apparent from the context g_{ij} - is a central quantity as it provides a link between information theory and geometry. The fisher metric is defined on a statistical manifold of parametrised probability distributions $p(x;\xi)$. It usually written in one of two standard forms (Amari and Nagaoka, 2000):

$$g_{ij}^{\texttt{Fisher}} = \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi\right] = -\mathbb{E}_\xi \left[\partial_i \partial_j \ell_\xi\right] \tag{4.5.5}$$

where $\partial_i = \frac{\partial}{\partial \xi^i}$ and $\ell_\xi = \log p(x;\xi)$. The second form is only correct when the p_ξ is normalised[11].

The Fisher metric is a Riemannian metric tensor. It is the general form metric that I will use in later sections to derive specific colour spaces.

Importantly, it was shown by Chenstov (1982) that the Fisher metric is *the unique* Riemannian metric that is invariant under (Fisher) sufficient statistics (see footnote[12]). Therefore the Fisher information is special: it is neither dependant on the specific coordinates used to parametrise it (it's Riemannian), nor is it dependant on any particular representation of probabilities (it's Fisher sufficient).

Every (smooth) h-divergence has, as it's second order term, some multiple of the Fisher information. So no matter what h-divergence one chooses, it is determined by the Fisher information at small distances. We have a quantity that is common to all risk based measures of colour difference.

[10] when one does not have a normalised probability distribution, one must multiply this by $\tau = \int_\chi p_\xi \mathrm{d}\mu(x)$.

[11] For the same reasons as the other form requires multiplication by τ.

[12] A sufficient statistic has a rather technical definition which can be easily found in a good statistics book. It is hard to describe in words without making the rather tautological claim that it is an information preserving transformation of a probability distribution. An important example, however, would be a deterministic transformation. A sufficient statistic is any transformation of a probability distribution that does not, in effect, add noise.

The Fisher Metric and the Variational Distance

The variational distance is not smooth at $u = 1$: we cannot derive the Fisher information by using a Taylor expansion. This problem can be overlooked by noting that all of the Bayes consistent loss functions (as discussed in section 4.2) that yield smooth h-divergences *do* lead to a multiple of the Fisher information. The Fisher information is common to all smooth h-divergences that are equivalent to the variational distance - this is in the sense that they lead to the same optimal classification. Put another way, where a metric tensor *can be* defined by an h-divergence, it is proportional to the Fisher information. All optimal, Bayes consistent classifications correspond to the Fisher metric, or leave a metric undefined. Or, in other words, where a metric is defined at all *the Fisher metric is common to the entire class decision rules that optimally reduce the error probability.*

4.5.2 Connection Coefficients

As well as demonstrating the uniqueness of the Fisher metric, Chenstov (1982) shows that invariance to sufficient statistics also lead to a definition of the connection coefficients. The connection he describes is known as the α-connection and has coefficients defined by:

$$\Gamma^{(\alpha)}_{ij,k} = \mathbb{E}_\xi \left[\left(\partial_i \partial_j \ell_\xi + \frac{1-\alpha}{2} \partial_i \ell_\xi \partial_j \ell_\xi \right) (\partial_k \ell_\xi) \right] \tag{4.5.6}$$

The α here is the parameter the defines a particular geodesic (as in 4.4.2). Each set of connection coefficients, of which there is one set for each α, describes a different way of drawing the optimal line between two points. To calculate them we take the observation of Amari and Nagaoka (2000) that the coefficients can be identified in the third order expansion of the divergence (see Appendix C.1) where we see that as a consequence

$$-\alpha = 3 + 2\frac{f'''(1)}{f''(1)} \tag{4.5.7}$$

This does not mean, however, that integrating along such a curve yields the h-divergence. Indeed, it only leads to it approximately (fourth order in $\Delta\xi$).

4.5.3 Riemannian Connections

As the h-divergences are symmetric, they behave in a nicer way than f-divergences in general. Their symmetry can be expressed as:

$$D^{(h)}(\,\xi \,||\, \rho\,) = D^{(h)}(\,\rho \,||\, \xi\,) \tag{4.5.8}$$

meaning that we can write (remembering $u = \frac{p_\rho}{p_\xi}$):

$$\begin{aligned} \forall p_\xi, p_\rho : \int_\chi p_\xi f(u) \mathrm{d}\mu(x) &= \int_\chi p_\xi u f\left(\tfrac{1}{u}\right) \mathrm{d}\mu(x) \\ f(u) &= u f\left(\tfrac{1}{u}\right) \end{aligned} \tag{4.5.9}$$

Taking the third order derivative we can see that $\alpha = 0$:

$$\begin{aligned} f'''(u) &= -3\frac{f''\left(\frac{1}{u}\right)}{u^4} - \frac{f'''\left(\frac{1}{u}\right)}{u^5} \\ 2f'''(1) &= -3f''(1) \\ \alpha &= -3 - 2\frac{f'''(1)}{f''(1)} = 0 \end{aligned} \tag{4.5.10}$$

This means that the affine connection coefficients associated with h-divergences are the Riemannian connection coefficients, α-geodesics with $\alpha = 0$:

$$\Gamma^{(0)}_{ij,k} = \mathbb{E}_\xi \left[\left(\partial_i \partial_j \ell_\xi + \tfrac{1}{2} \partial_i \ell_\xi \partial_j \ell_\xi \right) \left(\partial_k \ell_\xi \right) \right] \tag{4.5.11}$$

which can be seen to be the Riemannian connection by taking equation 4.4.9 and letting $g_{ij} = \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi \right]$ (this is a standard result, see Amari and Nagaoka, 2000, p33). Firstly:

$$\partial_k \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi \right] = \mathbb{E}_\xi \left[\partial_i \partial_k \ell_\xi \partial_j \ell_\xi \right] + \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \partial_k \ell_\xi \right] + \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi \right]$$

the last term of which is due to the expectation being taken over a distribution parametrised by ξ. We now have:

$$\begin{aligned} \Gamma^{(\text{Levi-Citiva})}_{ij,k} &= \tfrac{1}{2} \left(\partial_i \mathbb{E}_\xi \left[\partial_k \ell_\xi \partial_j \ell_\xi \right] + \partial_j \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_k \ell_\xi \right] - \partial_k \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi \right] \right) \\ &= \mathbb{E}_\xi \left[\partial_i \partial_j \ell_\xi \partial_k \ell_\xi \right] + \tfrac{1}{2} \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi \right] \\ &= \Gamma^{(0)}_{ij,k} \end{aligned} \tag{4.5.12}$$

This symmetry leads to a zero third order term in the expansion of the divergence in terms of $\Delta\xi$, meaning that the euclidian distance approximates h-divergences an oder better than non-symmetric divergences.

4.5.4 Variational Distance and Other Divergences

As I mentioned above, the variational distance, which directly yields the probability of classification error, cannot be directly related to the Riemannian arc-length by geometric considerations due to the non-differentiability of $f(u) = |1 - u|$ at $u = 1$. And, to do so in any other way requires knowing more than the value of the divergence.

Although there is not (necessarily[13]) a bijective relationship between the variational and smooth h-divergences, there are a number inequalities that relate the two (Kullback, 1966; Topsœ, 2000).

The most well known inequality is Pinsker's[14] inequality, refined by Fedotov et al. (2003) which relates variational distance V and the Kullback-Leibler divergence D_{KL}[15] (an f, but not h, -divergence).

$$D_{KL}(P||Q) \geq \tfrac{1}{2}V(P,Q)^2 \tag{4.5.13}$$

In many applications assumptions can be made that allow one to uniquely define a one-to-one relationship between the variational distance. For example, there is a sizable section of Wyszecki and Stiles (2000, p682) devoted to, in effect, making such a calculation assuming Gaussian variables (though this is never made explicit).

4.6 Proof of Concept

Before looking at how this theory can be put in a general setting, it is good to see an example that shows that it works. Here, I take a basis for a colour space as described in Vorobyev and Osorio (1998) as a starting point. The space is based upon Poisson statistics, so the example here uses Poisson distributions and their statistics also. This example also demonstrates a potentially useful trick. We need not calculate a metric directly, we can instead use small differences in parameters as measured by a divergence as an approximation to the metric.

Here I have formulated a model of spectral sensitivity (more specifically a $\Delta\lambda$ curve) using the KL-Divergence as a divergence that approximates the metric. This curve represents the ability for an observer to discriminate two monochromatic lights, with wavelengths λ and $\lambda + \Delta\lambda$: the value of $\Delta\lambda$ being that required for a particular adeptness in discrimination (in this case set to c). The definition of the quantities here follows exactly the same method as I will outline in part 5.4, except the calculation of KL-Divergence is used instead of the Fisher metric.

[13]It is clear that when two divergences share a geodesics that one will be representable as a function of the other, as they are both functions of the arc-length. When the geodesics are undefined, as in the case of the variational distance, it is difficult to formally make such comparisons. There is still a gap in the information theory literature with regard to this problem.

[14]Originally reported in Russian: M. S. Pinsker, Information and Information Stability of Random Variables and Processes. Moscow, U.S.S.R.: Izv. Akad. Nauk, 1960.

[15]The KL-divergence is defined by $f(u) = -\log(u)$ or $f(u) = u\log(u)$. In the original paper by Kullback and Leibler it is defined as $(1-u)\log u$ which *is* an h-divergence (Kullback and Leibler, 1951)

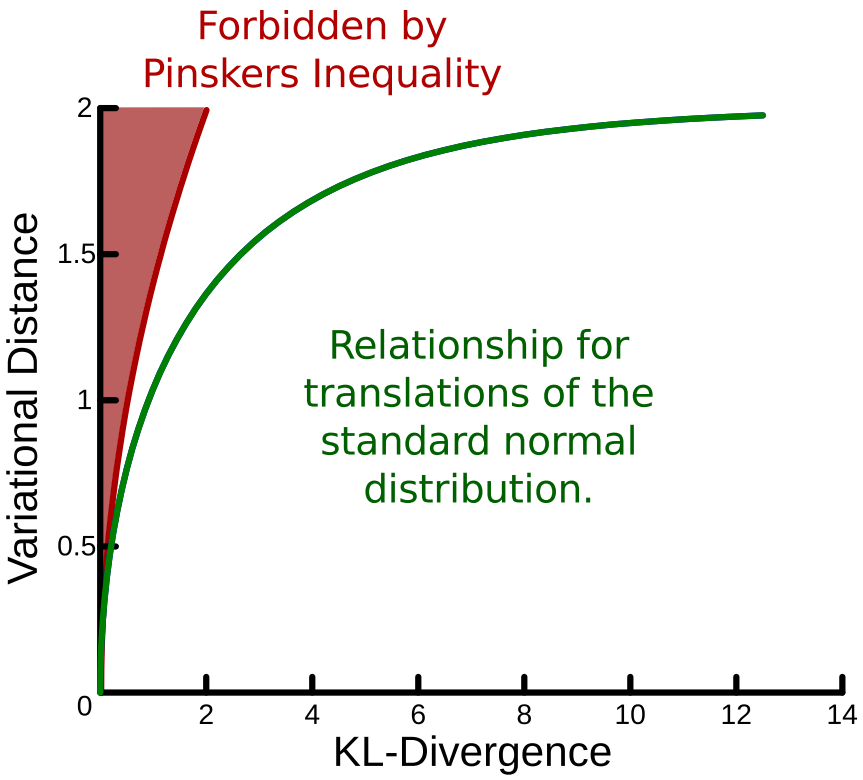

Figure 4.5: An example of the relationship between the Kullback-Leibler Divergence and the Variational Distance: The red area is forbidden for any distribution by Pinskers inequality. Pinskers inequality can be refined making the red area larger. For many model families, such as the one here: $p(x;\xi) = e^{-(x-\xi)^2/2}/\sqrt{2\pi}$, the relationship between D_{KL} and V is bijective. Note: the green line is the same for both $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$.

The approximation can be expressed as:

$$D^{(KL)}\left(\lambda \,||\, \lambda + \Delta\lambda\right) \approx g_{ij}(\xi)\frac{\partial\xi}{\partial\lambda}\frac{\partial\xi}{\partial\lambda}(\Delta\lambda)^2 = g(\lambda)(\Delta\lambda)^2 \tag{4.6.1}$$

and setting this to a constant discriminability (c) we have:

$$D^{(KL)}\left(\lambda \,||\, \lambda + \Delta\lambda\right) = c \tag{4.6.2}$$

which is then solved numerically for small, constant c to produce the $\Delta\lambda$ curves.

A comparison is shown in figure 4.6. The two curves differ very slightly in their loci in quantum-catch space, in Vorobyev and Osorio (1998) the spectral locus is taken with equal spectral power, then the luminance contribution is discarded, in the calculation I present here, the quantum catches of the monochromatic stimuli are forced to be within the same isoluminant plane by adjusting their power. Even given this difference, it is easy to see that approximations using divergences do agree well with a standard model in animal colour vision.

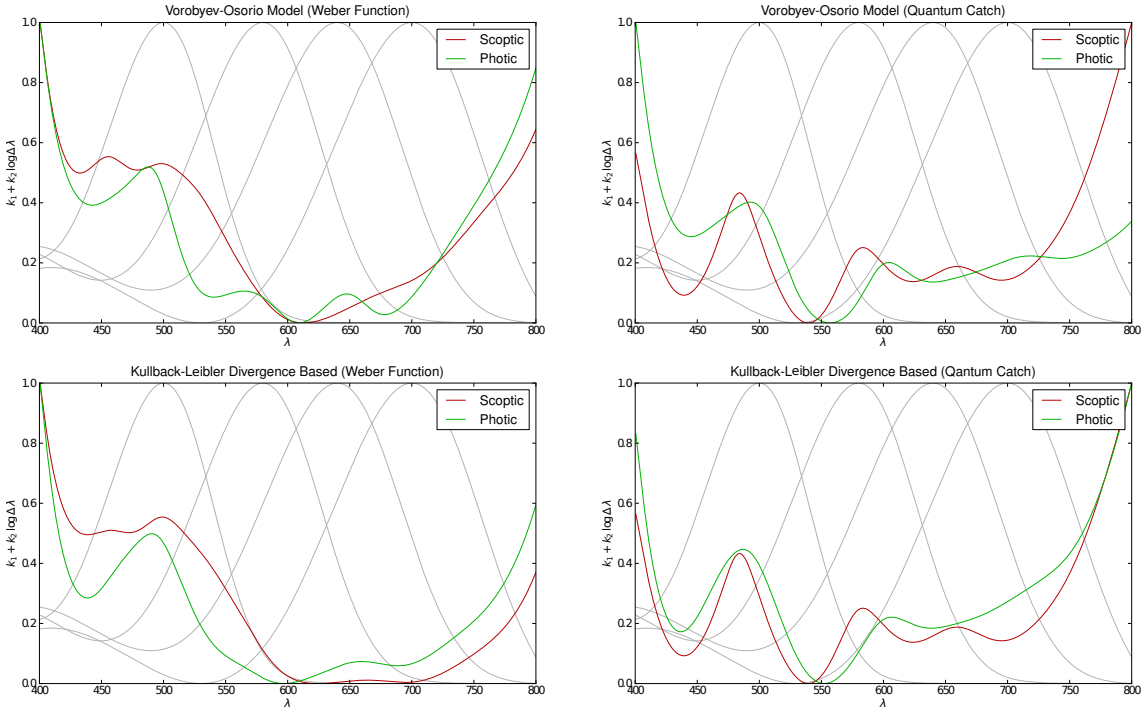

Figure 4.6: Comparison of the Vorobyev-OsorioVorobyev and Osorio (1998) model with the equivalent model based on the KL-divergence (a non approximate form of 5.4). The graphs show predicted wavelength discrimination ($\Delta\lambda$), assumed to be inversely proportional to the measure of discriminability between two neighbouring monochromatic spectra (delta functions centred at λ). The values of $k_{(i)}$ in the KL-divergence model are 0 for photic and 1000 for scoptic vision. The two models agree very well, considering the differences outlined in the body text - the difference between the extremities of the graphs is because of the difference in normalisation described. The grey lines describe the spectral dependence of each of the photoreceptor classes.

4.7 Extensions

I will now consider a number of extensions to the framework presented so far.

4.7.1 Asymmetric Loss and Non-Riemannian Geodesics

The theory presented so far is explicitly related to symmetric loss functions, as such losses are need for Bayes consistency. The measure of loss from erring with $y = 1$ has been the same as where $y = -1$. Indeed, this is why it was possible to write these losses as $\phi(y\gamma(x))$. However, in many settings, we may be interested in losses that are not symmetric. To use a well known example from colour theory: the loss associated with a train driver misidentifying a red light as green is much greater than the loss associated with mistaking

a green light for a red one.

Relaxing the symmetry condition requires us to write the risk function as a combination of two functions ϕ^+ and ϕ^- which have the same constraints as the symmetric ϕ-losses (convexity etc.):

$$\phi(\gamma(x), y) = \begin{cases} \phi^+(\beta), & y = 1 \\ \phi^-(-\beta), & y = -1 \end{cases} \tag{4.7.1}$$

This risk, like before, is given by - letting $\pi(x) = p(x, y = 1)$ and $\psi(x) = p(x, y = -1)$:

$$\begin{aligned} \mathcal{R}_\phi(\gamma) &= \mathbb{E}\left[\phi(\gamma(x), y)\right] \\ &= \int_\chi \left\{\phi^+(\gamma(x))\pi(x) + \phi^-(-\gamma(x))\psi(x)\right\} \mathrm{d}\mu(x) \\ &= \int_\chi \pi(x) \left\{\phi^+(\gamma(x)) + \phi^-(-\gamma(x))\frac{\psi(x)}{\pi(x)}\right\} \mathrm{d}\mu(x) \end{aligned} \tag{4.7.2}$$

and the optimisation with respect to γ yields an f-divergence:

$$\begin{aligned} \mathcal{R}^*_\phi &= \inf_\beta \int_\chi \pi(x) \left\{\phi^+(\beta) + \phi^-(-\beta)\frac{\psi(x)}{\pi(x)}\right\} \mathrm{d}\mu(x) \\ &= \int_\chi \pi(x) \inf_\beta \left\{\phi^+(\beta) + \phi^-(-\beta)\frac{\psi(x)}{\pi(x)}\right\} \mathrm{d}\mu(x) \\ &= -\int_\chi \pi(x) f(u) \mathrm{d}\mu(x) \end{aligned} \tag{4.7.3}$$

so

$$f(u) = -\inf_\beta \left\{\phi^+(\beta) + u\phi^-(-\beta)\right\} \tag{4.7.4}$$

No longer does $f(u) = uf(1/u)$ and this means that the relationships that ensured the Riemannian connection no longer holds: $\alpha = 0$ is not by necessity true. Asymmetric losses imply asymmetric divergences and non-Riemannian connection coefficients.

Example of Asymmetric Divergence from Asymmetric Loss

If we take the exponential loss functions $\phi(\beta) = e^{-\beta}$ as a starting point, using it to define an asymmetric loss so that:

$$\phi^+(\beta) = \phi(\beta) \tag{4.7.5}$$

$$\left(\phi^-(\beta)\right)^k = \phi(\beta) \tag{4.7.6}$$

k (which is ≥ 0) can be thought of as putting complete emphasis on ϕ^+ when it tends to zero and complete emphasis on ϕ^- as it tends to infinity; at 1 it is symmetric and Bayes consistent. This defines a divergence such that:

$$f(u) = (ku)^{\frac{1}{k+1}} + u(ku)^{\frac{-k}{k+1}} \tag{4.7.7}$$

which is geometrised so that:

$$\begin{aligned} h(u) &= \frac{1}{k}\left((1+k)(ku)^{\frac{1}{1+k}} - k^{\frac{1}{1+k}}(k+u)\right) \\ g_{ij} &= \frac{k^{\frac{1}{k+1}}}{k+1} g_{ij}^{\texttt{Fisher}} \\ \alpha &= \frac{k-1}{k+1} \end{aligned} \tag{4.7.8}$$

When $k = 1$ this is the Hellinger distance. The geodesics that are defined by this divergence with other values of k range between $\alpha = \pm 1$ in the limits of $k \to \infty$ and $k \to 0$. At these limits, the geodesics are the same as those given by the Kullback-Leibler divergence (though such limits are behaviourally irrelevant, as it corresponds to infinite risk for one alternative).

A Testable Hypothesis

The difference in geodesics has testable outcomes. Take the following generalisation experiment; As conditioning, we use two colours for which rewards are given upon being selected - this is used to train the subject so that they develop a preference for these two colours.[16] Unconditioned stimuli of different colours lying on a line half way between the two conditioning stimuli as measured by the various α-geodesic distances are then presented and the preference for each recorded.

The theory presented here would suggest that in the case of equal rewards the preferred colour would, approximately, lie on the geodesic for which $\alpha = 0$, whereas, in the case of unequal rewards (or punishments) a different geodesic would be expected. Let us now look in more detail at a simple experimental paradigm that has the potential to illuminate this phenomenon.

The first step is to select a diamond in colour space, $ABCD$, so that each vertex corresponds to a colour that will be used in the experiment. The lengths of the edges are chosen so that some information measure with $\alpha = 0$ is equal for all edges - this corresponds to a euclidean diamond in a 'perceptually uniform' space. Furthermore, we also require that under other divergences, where $\alpha \neq 0$, that $D^{(f)}(B \,||\, A) = D^{(f)}(B \,||\, C)$, $D^{(f)}(A \,||\, B) = D^{(f)}(C \,||\, B)$, $D^{(f)}(D \,||\, A) = D^{(f)}(D \,||\, C)$ and $D^{(f)}(A \,||\, D) = D^{(f)}(C \,||\, D)$. In other words, A and C should be both the same distance from B and both the same distance from D. This need not be exact, but the closer we get to this, the better the result.

[16] Though we could do the reverse and choose colours that are preferred.

The variation in risk must in some sense relate to the uncertainty. A simple way of doing this, would to be to put different limits to the response time for different stimuli. Even if there are prior (i.e. non-equal) risks associated with each stimulus, we can account for their contribution by exaggerating the difference.

As an example, I have chosen a simple space which is determined by Poisson statistics (and uniform prior probabilities) so that the metric determining statistics are given by Poisson distributions with a rate parameters given by the quantum catch (see section 5.4). To do this I take a standard euclidean square in the coordinate system with coordinates $\xi^i = \sqrt{q^i}$, where q^i is the quantum catch. I choose a square such that $\sum_i \xi^i = 1$, so as to mimic chromaticity space. This example uses trichromatic colour space.

Point	ξ^1	ξ^2	ξ^3	q^1	q^2	q^3
A	0.663953	0.663953	0.404145	0.440833	0.440833	0.163333
B	0.727350	0.427350	0.577350	0.529038	0.182628	0.333333
C	0.490748	0.490748	0.750555	0.240833	0.240833	0.563333
D	0.427350	0.727350	0.577350	0.182628	0.529038	0.333333

Table 4.2: Coordinates for colours in this example experimental setup.

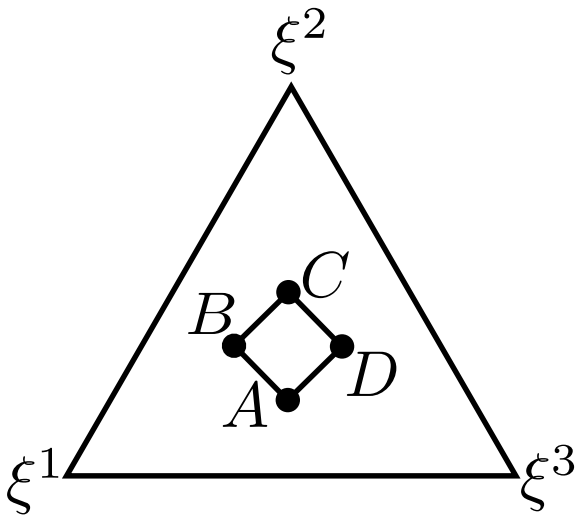

Figure 4.7: Diagram of the choice of points in section 4.7.1, representing to those in table 4.2. The triangle represents a chromaticity type space in ξ.

For the experiment, using an unbiased or unbiased set of rewards we train the subject to colours specified by B and D, then as a test we look at the preference for the colours A or C. When the rewards are equal we should expect no preference for either A or C, when

they differ, we should expect a bias. We can see this in table 4.3. The exact quantification of the distances is impossible as we do not know exactly which risk function is appropriate, but we can look at the qualitative change. We can think of the different f-divergences as all being monotonic functions of some natural α-geodesic distance (such as the α-divergences of Amari and Nagaoka (2000)), thus, if for some divergence with a given α, one distance is greater than another, this will also be the case for any other divergence with the same α. The changing of α in this experiment, is in effect moving the two training points closer or further away from one of the two test points (and a such requires a corresponding asymmetry to exist in the relationship between the squares coordinates and geometric structure of the statistical manifold). This is represented diagrammatically in figure 4.8.

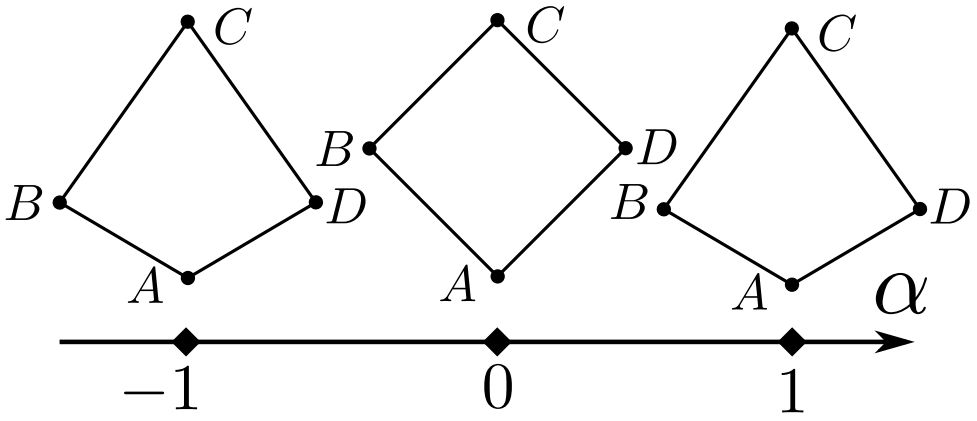

Figure 4.8: The 'apparent' geometry of the stimuli for different α (and thus, relative rewards in the training scenario). The divergences are measured either rightwards ($\overrightarrow{AD}$, $\overrightarrow{BA}$, $\overrightarrow{BC}$ and $\overrightarrow{CD}$) or leftwards ($\overrightarrow{AB}$, $\overrightarrow{CB}$, $\overrightarrow{DA}$ and $\overrightarrow{DC}$), or equivalently in one direction with $\pm\alpha$ (see Amari and Nagaoka, 2000, 3.1 - The Duality of Connections). Assuming that we are measuring divergences in a chosen direction, as we change $|\alpha|$, B and D move closer to either A or C, measuring in the other direction, the converse happens.

4.7.2 Different Prior Probabilities

I shall end the substantive part of this section by showing how the theory presented above can be extended to non-normalised probabilities. Up to now I have mostly considered measurements on conditional probabilities by assuming a maximum entropy prior distribution over the actual class $\mathcal{Y}$. The theory can be easily extended to prior class distributions that are not maximum entropy; it has been left to the end as it makes no difference to the previous arguments except for making formulae much harder to read. This is an application of the quite general procedure of denormalisation as outlined by Amari and

Choice	$\alpha = -1$	$\alpha = 0$	$\alpha = 1$
A	0.47498	0.50000	0.52495
C	0.52495	0.50000	0.47498

Table 4.3: Relative distances of A and C from B and D. The values in the table are given by $\frac{D^{(f)}(\,B\,||\,X\,)}{D^{(f)}(\,B\,||\,X\,)+D^{(f)}(\,X\,||\,D\,)}$, where X is either A or C. I have used the canonical divergences of Amari and Nagaoka (2000) as $D^{(f)}\,(\,\cdot\,||\,\cdot\,)$, these are the two Kullback-Leibler divergences for $\alpha = \pm 1$, and the Hellinger distance for $\alpha = 0$. Different choices would give different values but the trend would remain; the fraction either increasing or decreasing with α (depending on the choice of A or C) and with it staying constant at 0.5 for all $\alpha = 0$.

Nagaoka (2000). Denormalisation allows us to not only consider a manifold of probability distributions such that

$$\int_{\mathcal{X}} p(x)\mathrm{d}\mu(x) = 1 \tag{4.7.9}$$

but any value $\tau > 0$:

$$\int_{\mathcal{X}} p(x)\mathrm{d}\mu(x) = \tau \tag{4.7.10}$$

Using the same definition as equation 4.3.3, one can write a general f-divergence as:

$$\begin{aligned} D^{(f)}\,(\,y_1\,||\,y_2\,) &= \int_{\mathcal{X}} p(x,y_1) f\left(\frac{p(x,y_2)}{p(x,y_1)}\right)\mathrm{d}\mu(x) &(4.7.11)\\ &= \int_{\mathcal{X}} p(x|y_1)p(y_1) f\left(\frac{p(x|y_2)p(y_2)}{p(x|y_1)p(y_1)}\right)\mathrm{d}\mu(x) &(4.7.12)\end{aligned}$$

From here, we make the identifications:

$$\begin{aligned} p(x|y_1) &= p(x;\xi_1) &(4.7.13)\\ p(x|y_2) &= p(x;\xi_2) &(4.7.14)\\ p(y_1) &= \tau &(4.7.15)\\ p(y_2) &= 1-\tau &(4.7.16)\end{aligned}$$

but consider the more general case of divergences between (ξ_1,τ_1) and (ξ_2,τ_2) with the above special case being where $\tau_1 = \tau$ and $\tau_2 = 1-\tau$. This means that we have, letting $\rho_i = (\xi_i,\tau_i)$:

$$p(x,y_1) = \tau_1 p(x;\xi_1) = p(x;\rho_1) \tag{4.7.17}$$

$$p(x,y_2) = (1-\tau_1)p(x;\xi_2) = p(x;\rho_2) \tag{4.7.18}$$

so that in general we have f-divergences of the form:

$$D^{(f)}\left(\,\rho_1\,||\,\rho_2\,\right) = \int_{\mathcal{X}} p(x;\rho_1) f\left(\frac{p(x;\rho_2)}{p(x;\rho_1)}\right) \tag{4.7.19}$$

Defining the a new index τ which corresponds to partial derivatives as:

$$\partial_\tau = \frac{\partial}{\partial \tau} \tag{4.7.20}$$

The Fisher metric is extended to $\tilde{g}$ by:

$$\tilde{g}_{ij} = \tau g_{ij}, \;\; \tilde{g}_{i\tau} = 0, \;\; \tilde{g}_{\tau\tau} = \frac{1}{\tau} \tag{4.7.21}$$

or more graphically, for a parameter vector $(\xi^1 \cdots \xi^N, \tau)$:

$$\tilde{g} = \left(\begin{array}{ccc|c} & \vdots & & \vdots \\ \cdots & \tau g_{ij} & \cdots & \mathbf{0}^N \\ & \vdots & & \vdots \\ \hline \cdots & \mathbf{0}_N & \cdots & \tau^{-1} \end{array}\right) \tag{4.7.22}$$

Often we are concerned with a space with fixed prior probabilities - the sub-manifold with $\tau = q(\xi)$. i.e. where $q(\xi)$ is the prior probability at any ξ. This corresponds to cases where $q(\xi)$ is determined natural scene statistics or, if one wishes, other more contextualised distributions. It is fairly straight forward to calculate the metric on this sub-manifold, it is simply the value of $\bar{g}$ which yields a solution to:

$$\bar{g}_{ij} \mathrm{d}\xi^i \mathrm{d}\xi^j = \left(\tau g_{ij} + \frac{1}{\tau}(\partial_i \tau)(\partial_j \tau)\right) \mathrm{d}\xi^i \mathrm{d}\xi^j \tag{4.7.23}$$

$$\bar{g}_{ij} = \tau \left(g_{ij} + (\partial_i \log \tau)(\partial_j \log \tau)\right) \tag{4.7.24}$$

I shall not go into any more details here, except to show that not using prior probabilities (instead, just normalised and parametrised distributions) is equivalent to the uniform (maximum entropy) prior over all ξ. If we want a prior distribution that yields the same geometry as the conditional geometry (up to a multiplicative constant: $k\bar{g} = g$ for some $k > 0$), then we have the requirement:

$$k g_{ij} = \tau \left(g_{ij} + (\partial_i \log \tau)(\partial_j \log \tau)\right) \tag{4.7.25}$$

The easiest way of solving this is realising the second part of the sum must be zero:

$$\partial_i \log \tau = 0 \tag{4.7.26}$$

$$\tau = \textit{constant} \tag{4.7.27}$$

which is clearly a solution to the equation in general and corresponds to a uniform prior.

4.8 Summary of Framework

The theoretical framework presented above shows how colour spaces can be generated from the starting point of risk as found in machine learning literature. It will be seen in the next chapter that this theory leads to familiar colour spaces.

I have shown how a requirement for *Bayes consistency* of the *binary classifier* responsible for discriminating colours yields a whole *class of distance measures* - the h-divergences. These are symmetric. When it is possible to define a Riemannian metric from them, it is a multiple of the *Fisher metric*. All of them yield *the same geometry*. Whilst it is true that divergences in general do not give Riemannian geodesics, the class of h-divergences does. The curves along which h-divergences are minimised are all *Riemannian geodesics*. If we relax the constraint of symmetric loss functions the geodesics are potentially different. Non-Riemannian α-geodesics correspond to non-Bayes-consistent binary classifiers and we see that the weighting of the classifier towards a particular choice changes the geodesic and that the effect of this can potentially be observed experimentally.

For all risks I have mentioned there is a corresponding divergence, and as demonstrated in appendix C.1 and by the theorem of Chenstov (1982) this means that the Fisher metric is the *only Riemannian metric* to use for colour theory (and perceptual theory in general!) if one requires consistency with this model of risk. Such measures constitute a large class of global metrics that can all be justified as measures of long range colour difference - this reflects the difficulties that exist in deriving such quantities, in fact, to my knowledge no other theoretically justified derivations of such measures exist.

With the addition of the α-connection to the geometry of colour we also have a generalised notion of perceptual uniformity, with a different type of uniformity for each value of $|\alpha|$. The transformation of coordinates so that divergences become straight (euclidean) lines is not uniquely defined, but connected to the value of α. Making a space perceptually uniform in this sense applies to non-infinitesimal distances must therefore take this into account (using α-affinity, see Amari and Nagaoka, 2000). In biological systems the risks and rewards involved with making a decision are rarely symmetric. The results here are of great importance in such cases. However, in the contrived situation of an experiment where there is no bias in the risks or rewards involved, we can expect α to be zero and the transformation to perceptual uniformity to be the one that makes the Fisher metric proportional to the identity matrix.

Chapter 5

Local Models of Colour Vision

"In nature we never see anything isolated, but everything in connection with something else which is before it, beside it, under it and over it."

Johann Wolfgang von Goethe, 1749-1832

In this section I will review some existing models of colour that use differential geometric methods, formulating them in the framework of the previous section so that I may highlight some of the theoretical assumptions that they suggest.

Although under certain regimes photoreceptor responses are linear (de Ruyter van Steveninick and Laughlin, 1996), this is not always the case (Alleysson and Hérault, 1997; French et al., 1993). A commonly cited example of photoreceptor non-linearity is the Bezold-Brücke effect, where lights that are bright relative to the state of adaptation desaturate (Backhaus, 1992). There is a less obvious, but ubiquitous source of non-linearity: where there is non-Gaussian noise. This is so because in general we can find a non-linear transformation of the support such that the noise becomes Gaussian. Similarly, we can do the inverse; turn Gaussian noise on a non-linear coordinate[1] into non-Gaussian noise on a linear coordinate. This is where the coordinate invariant information geometric measures find their true place.

Here I investigate non-linearities arising from various non-Gaussian noise regimes and the connection between various existing colour spaces from the perspective of the framework I developed. I will provide the Fisher metric and (where possible) a transformation of the quantum catch coordinates that makes the space 'perceptually uniform'.[2]

Perceptually uniform in this case does not describe the experience of the subject, but simply to the affine nature of the geometry. Such correspondance could only be achieved

[1] By this I mean a coordinate non-linearly related to the transmitted signal - such as in Shannon (1948).

[2] The transformation $\rho(\xi)$ makes the space perceptually uniform if $g_{ij}(\rho) = k\delta_{ij}$.

if the model were perfect and the risk was entirely captured by the Euclidean distance. Perceptual uniformity in this case is simply a short hand to describe a cannonical form of a coordinate system which if derived from a good model would have rotationally symmetric and equal sized ellipsoids describing the ability to descriminate.

Notation

For this section I will adopt the convention that the coordinates $\xi \in \Xi$ refer to those of the quantum catch space - previously written q in 1.3.1. I use $\rho \in R$ to represent other coordinates systems. The support coordinates $x \in \mathcal{X}$, as before, parametrise the space whose elements are used to produce a classification.

5.1 Gaussian Noise

The most immediately apparent application of information theoretic approaches is the case where Gaussian noise is 'added to a signal', much like the cases used by (Shannon, 1948) to derive information measures based on signal to noise ratios. This simple model is the point of departure for the rest of the chapter, in which the models will become increasing detailed. However, it can be shown that this situation is broadly applicable and that a similar result holds for all members of the exponential family whose parametrisation define only the mean and leave all other moments constant[3]. Although this model is rather simplistic, it is important as the Gaussian distribution plays a central role in information theory and the metric space it induces is affine. From now on I will make use of some the notational devices in appendices A.3.1, A.3.2 and A.3.3.

The model of the colour system of an n-chromat is simply taken to be the statistical manifold described by the probability density functions representing independent Gaussian noise around a mean value of ξ:

$$p(x;\xi) = \prod_{k=1}^{n} \frac{1}{\sigma^k\sqrt{2\pi}} \exp\left(-(\xi^k - x^k)^2/2(\sigma^k)^2\right) \tag{5.1.1}$$

Each variate in $\mathcal{X}$ is independent so we can immediately see that $g_{ij}(\xi) = 0$ when $i \neq j$. For the same reason it is possible to treat each case where $i = j$ separately - letting us

[3]See Amari and Nagaoka (2000)

drop the indices for now. I will use ∂ to denote $\frac{\partial}{\partial \xi}$.

$$\begin{aligned} \ell(x;\xi) &= -\frac{(\xi - x)^2}{2\sigma^2} + constant \\ \partial\ell(x;\xi) &= -\frac{(\xi - x)}{\sigma^2} \\ \mathbb{E}\left[\partial\ell(x;\xi)\partial\ell(x;\xi)\right] &= \int_{-\infty}^{\infty} \frac{(\xi - x)^2}{\sigma^4} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\xi-x)^2}{2\sigma^2}} \mathrm{d}x \end{aligned} \tag{5.1.2}$$

which can be evaluated (using parts) to:

$$\mathbb{E}\left[\partial\ell(x;\xi)\partial\ell(x;\xi)\right] = \frac{1}{\sigma^2} \implies g_{ii} = \frac{1}{\sigma_{(i)}^2} \tag{5.1.3}$$

so

$$g_{ij} = \frac{\delta_{ij}^k}{\left(\sigma^k\right)^2} \tag{5.1.4}$$

The metric in this case is equal to the inverse of the covariance matrix of that would describe the multivariate Gaussian in equation 5.1.1 - as the coordinates of $\mathcal{X}$ are independent such a covariance matrix would be diagonal. It is not particularly hard to derive the case where Gaussian noise is correlated between the photoreceptors, in which case, the metric is given by the inverse of a non-diagonal covariance matrix. This justifies the commonly used covariance matrix derivation (Vorobyev and Osorio, 1998; Wyszecki and Stiles, 2000), but also shows that it can only be taken to be realistic when the assumption of independent additive Gaussian (up to change in measure) noise is applicable.

5.2 Weber's Law: Helmholtz's and Stiles' Spaces

The colour model of von Helmholtz (1896)[4] represents the earliest non-trivial colour space model. This space has many justifications, I wish to use it here as a means of idealising two extremes of explanation - one where the metric arises deterministically, the other where it arises statistically. The examination of this space provides an entry point to a discussion about the correct derivation of colour space metrics.

The Helmholtz space has the metric:

$$g_{ij} = \frac{\delta_{ij}^{kl}}{\left(\tau^l \xi^k\right)^2} \tag{5.2.1}$$

which is often justified as the application of Weber's law to each photoreceptor channel. This can be seen by choosing a straight line $\gamma(t)$:

$$\gamma(t) = tz^i + c^i \tag{5.2.2}$$

[4]See also Wyszecki and Stiles (2000, p658)

for $z^i \in \mathbb{R}_+$ with the constraint that z is a unit vector ($z_i z^i = 1$). Then:

$$\begin{aligned} \mathrm{d}s^2 &= g_{ij} \frac{\partial \gamma^i(t)}{\partial t} \frac{\partial \gamma^j(t)}{\partial t} \mathrm{d}t^2 \\ &= \delta^k_{ij} \frac{z^i z^j}{\left(\tau_{(k)} \xi^k\right)^2} \mathrm{d}t^2 \end{aligned} \tag{5.2.3}$$

so that when a line which only changes in one photoreceptor coordinate, m, i.e. $z^i = z^m = 1$ for some $i = m$ then:

$$\mathrm{d}s = \frac{\mathrm{d}t}{\tau_{(m)} \xi^m} \tag{5.2.4}$$

To get Weber's law, we consider a linear approximation:

$$\Delta s = \delta^{ijm} \frac{\Delta t}{\tau^i \xi^j} \tag{5.2.5}$$

If we identify $\frac{\Delta t}{\Delta s}$ the apparent change with respect to the actual change - i.e. a small change in the coordinates: $\Delta\xi$ we get, after dropping the indices:

$$\frac{\Delta \xi}{\xi} = \tau \quad (= \omega) \tag{5.2.6}$$

where ω is known as the Weber fraction. We say 'the discriminability is proportional to the intensity'. We should note that this doesn't apply to changes in any directions other than along the coordinate axes i.e. $z = (1, 0, 0, ...)$, $(0, 1, 0, ...)$ etc. Weber's law cannot be taken to apply both at the photoreceptor level and in the colour space at large.

But how is this justified mechanistically? Is it justified at all? There have been many attempts to justify Weber's law mechanistically (Cope, 1976; Deco and Rolls, 2006; Masin et al., 2009; Shen and Jung, 2006), the problem it seems, is not that it's hard to derive, but the complete opposite - it is too easy - there are so many different ways that it can arise that it is impossible to say that it is any particular one without further argument.

I will now provide two interpretations of this metric, one statistical, the other deterministic.

5.2.1 Derivation as a Statistical Phenomenon

The firing of sensory neurons can often be described as a Poisson processes, neurons that can be described like this are known as Poisson neurons. One could consider a population of N Poisson neurons whose mean inter-spike interval[5] is proportional to signal intensity as defined by the quantum catch ξ (spike time encoding, see e.g. Sanderson et al., 1973).

[5]The inter-spike interval is the time between action potentials. The distribution is typical of sensory neurons.

This situation implies that each of the population's inter-spike intervals, x^i, are distributed according to a gamma distribution.

$$Pr(x^i = x^i_\alpha) = \frac{(x^i_\alpha)^{N+1} e^{-x^i_\alpha/\tau\xi^i}}{\Gamma(N)(\tau\xi^i)^N} \tag{5.2.7}$$

With some work it is possible to derive equation 5.2.1 as the Fisher metric for this distribution (see Appendix D.2). It results in an equation of the form:

$$g_{ij} = \frac{N\delta^k_{ij}}{(\tau\xi^k)^2} \tag{5.2.8}$$

which is the Helmholtz metric.

5.2.2 Interpretation as a Deterministic Phenomenon

It is possible to assume that photoreceptors respond logarithmically to the amount of light incident upon them (see e.g. Koshitaka et al., 2008). Using this we can calculate the transformation from the quantum catch type space (parametrised by ξ) to a photoreceptor type space (parametrised by $\rho = [\rho^a]$). As the channels are independent we can do the calculations without indices, as in the first section:

$$\rho = \log\xi \quad \text{so} \quad \frac{\partial\xi}{\partial\rho} = \xi \tag{5.2.9}$$

We now get the metric in this coordinate system that corresponds to the Helmholtz metric in the quantum catch coordinates (this is to show that when we use this parametrisation the metric is affine).

$$g_{ab}(\rho) = g_{ij}(\xi)\frac{\partial\xi^i}{\partial\rho^a}\frac{\partial\xi^j}{\partial\rho^b} = \frac{\delta^k_{ij}}{\tau^2_{(k)}(\xi^k)^2}(\delta^i_a\xi^i)(\delta^j_b\xi^j) = \frac{\delta^k_{ab}}{\tau^2_{(k)}} \tag{5.2.10}$$

We see that when parametrised by the logarithm of the quantum catch the Helmholtz space is affine. The affine (Gaussian based) metric on logarithmic quantum catch space is the same as the Helmholtz metric on (linear) quantum catch space.[6]

5.2.3 Two Explanations, One Result

It seems that without any extra argument neither of these is justified more than the other. Indeed, attempting to reverse engineer a single result to find a complex mechanism is not very sensible in the first place. However, there may be additional reasons to accept or reject a given way of arriving at a particular result when many alternatives exist. The

[6]The original coordinate system in which Helmholtz defined the metric was this (linear in quantum catch).

following section argues that when addressing psychophysical metrics we should reject a deterministic model in favour of a statistical model. More precisely, I argue that deterministic models such as the one above are actually statistical models in disguise.

5.3 The Fallacy of Metrics Based on Noiseless Physiological Transformations

I have described two potential explanations for the Helmholtz metric that seem to be at odds with each other. One is based on a physiologically motivated, deterministic transformation, the other on the addition of physiologically motivated, non-Gaussian noise. I would like to briefly discuss how deterministic transformations should be applied and how noise based spaces are related. I will then argue that in actuality a physiological space is a noise based space with tacit assumptions about the nature of the noise. To illustrate this I will begin by beating on a straw man - the claim that "you can have a colour metric without noise".

Consider a one dimensional system completely without noise. The value of a signal s is transformed by a one-to-one function f into a probability distribution p on support $\mathcal{X}$ with elements x and parametrised by s. This is completely general, it captures all possible (continuous) deterministic relationships between the 'input' s and the 'output' x:

$$p(x;s) = \delta(x - f(s)) \tag{5.3.1}$$

where δ, unlike elsewhere in this document, is the Dirac delta[7]. As the Dirac delta is discontinuous, let us instead consider the distribution:

$$\bar{p}(x;s) = \frac{\exp\left(-\frac{(x-f(s))^2}{2k^2}\right)}{k\sqrt{2\pi}} \tag{5.3.2}$$

which converges to the distribution in equation 5.3.1 in the limit $k \to 0$:

$$\lim_{k\to 0} \bar{p}(x;s) = p(x;s) \tag{5.3.3}$$

To calculate the Fisher metric we need to assume that it can take values in the extended real numbers $\bar{\mathbb{R}}$ with their usual algebraic definitions[8]. The Fisher metric for the $\bar{p}$ system

[7] This is defined so that $\delta(x) = 0$ if $x \neq 0$ and $\int_{\mathcal{X}} \delta(x) = 1$.

[8] The extend reals ($\bar{\mathbb{R}}$) are the real numbers with the addition of points at $\pm\infty$ so that $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ In this case these additional numbers obey, for all $x \in \bar{\mathbb{R}}, y \in \mathbb{R}$: $-(+\infty) = -\infty$; $+\infty > -\infty$; $+\infty > y$; $-\infty < y$; $x + (+\infty) = +\infty$, for $x \neq -\infty$; $x - (+\infty) = -\infty$ for $x \neq +\infty$; $x \times (\pm\infty) = \pm\infty$ for $x > 0$; $x \times (\pm\infty) = \mp\infty$ for $x < 0$; $y/\pm\infty = 0$; $\pm\infty/y = \pm\infty$ for $y > 0$ and $\pm\infty/y = \mp\infty$ for $y < 0$.

is:

$$g_{ij} = \frac{\delta_{ij}}{k^2} \tag{5.3.4}$$

so that in the limit where $\bar{p} \to p$ we have:

$$\lim_{k\to 0} g_{ij} = \lim_{k\to 0} \delta_{ij}\frac{1}{k^2} = \begin{cases} +\infty, & i = j \\ 0, & \textit{otherwise} \end{cases} \tag{5.3.5}$$

and similarly for the linearised distance between two points arc-length Δs we have:

$$\Delta s^2 = g_{ij}\Delta\xi^i\Delta\xi^j = \frac{|\Delta\xi|^2}{k^2} \qquad \text{so} \qquad \Delta s = \frac{|\Delta\xi|}{k} \tag{5.3.6}$$

then

$$\begin{aligned} \lim_{k\to 0} \Delta s &= \begin{cases} \lim_{k\to 0} \frac{|\Delta\xi|}{k}, & |\Delta\xi| > 0 \\ 0, & |\Delta\xi| = 0 \end{cases} \\ &= \begin{cases} +\infty, & |\Delta\xi| > 0 \\ 0, & |\Delta\xi| = 0 \end{cases} \end{aligned} \tag{5.3.7}$$

Here, every signal is perfectly discriminable from every other. By it, everything is either exactly the same or completely different. There is in fact a metric which has this property, it is called the discrete metric (Sutherland, 1975) and the distance between two stimuli x and y is given by:

$$d(x,y) = \begin{cases} 0, & x = y \\ k, & \textit{otherwise} \end{cases} \qquad \text{where } k \text{ is a strictly positive constant} \tag{5.3.8}$$

This metric could be described as categorical. Each point is a unique in every possible way. There are no degrees of sameness, just 'identical' and 'different'. In fact, the topology it induces is completely different from the other metrics I have considered (Sutherland, 1975) - it is so perverse that the (topological) dimension of the colour space is no longer the number of photoreceptor classes, but is instead 0. The space is a sea of isolated points; there is no concept of a point being 'near' another; *there is no geometry at all!*.

While we have got a metric from this transformation, it is not Riemannian, nor does it even correspond to a more general differential manifold.[9] On this basis we can quickly see that it does not correspond to any known colour model. Nor does it correspond to our sensibilities about biological systems. A sensory system with perfect fidelity is unheard of.

[9]Such as those described by multidimensional scaling. This technique is widely applied, in domains including colour (Backhaus et al., 1984), sound (Amézquita et al., 2011) and electrosensation (Von der Emde and Ronacher, 1994).

Nor does there exist an organism capable of a myriad of contingent actions. This would require an oracle of a scale that only exists in works of fiction (Borges, 1941).

A noiseless system is both unintuitive and leads to unrealistic metrics. Conversely, as the discrete metric of equation 5.3.8 is a consequence of all deterministic mappings[10], when there is geometry other than the discrete one we *must* infer there is indeterminacy underlying it.

A physiological result that suggests an input is deterministically transformed in some way tells us absolutely nothing about perceptual metrics without the addition of further assumptions!

Later in this chapter I will give an example of a case where this was not recognised and then go on to derive a similar result in the correct way. But for now I will continue my review of established colour metrics as they lead to this space incrementally.

5.4 Schrödinger's, and Vos and Walveren's Spaces

Spaces such as that of Schrödinger (1920) and Vos and Walraven (1972a,b), as well as that of Vorobyev and Osorio (1998) can be thought of as being based on Poisson statistics and in their simplest form upon the Poisson distribution parametrised by a linear function of the quantum catch. This has so far been calculated by substituting the variance of a Poisson distribution into the euclidean metric equivalent to the inverse covariance matrix. In this case this method does not yield a different result from the Fisher metric (though this does not hold more generally). To derive these three Poisson based spaces we calculate a metric based on two Poisson processes:

1. Photon noise: the number of photoisomerisation events happens at rate ξ.
2. Dark noise: the photopigment molecules spontaneously isomerise at rate k. This is a constant parameter for these models.

In the case of (Schrödinger, 1920), k is set to zero. Explicitly, in the space of Vorobyev and Osorio (1998), and implicitly in the Stiles space, photoreceptor density is taken into account: the incident light is summed over a given area of the retina and is thus dependent on the photoreceptor density d. For a particular area, A we have[11]:

$$X \sim \textsc{Poisson}\left(dA(\xi + k)\right) \tag{5.4.1}$$

[10] The methodology here only applies to continuous mappings, but a derivation for discontinuous f should be possible using the foundation provided by Lebanon (2005).

[11] We can consider multiple photoreceptors by the rate additivity of the Poisson distribution.

so that, more generally in the case of multiple receptor types the Fisher metric is given by (see Appendix D.1):

$$g_{ij} = \frac{\delta^l_{ij}}{Ad_{(l)}(\xi^l + k_{(l)})} \tag{5.4.2}$$

In the case of scotopic (night-time) vision, $k_{(l)} \gg \xi^l$ and:

$$g_{ij} \approx \frac{\delta^l_{ij}}{Ad_{(l)}k_{(l)}} \tag{5.4.3}$$

which defines a euclidean space: The metric has constant diagonal elements, much as in the Gaussian case presented earlier.

In photopic (day-time) conditions $k_{(l)} \ll \xi^l$:

$$g_{ij} \approx \frac{\delta^l_{ij}}{Ad_{(l)}\xi^l} \tag{5.4.4}$$

which, unlike the scotopic space is not affine in the quantum catch coordinates - it is not automatically perceptually uniform. When this is taken as an equality is equivalent to the space of Schrödinger (1920) at fixed luminance. It is also interesting to note that if Δs^2 is taken to be an apparent distance, then this space obeys Fechner's law in a non-coordinate determinant manner (unlike the Helmholtz space). As I described in chapter 4, divergences can be thought of as squared arc-lengths as $D^{(f)}\left(\,\xi \,||\, \xi + \Delta\xi\,\right) \approx \Delta\xi_i \Delta\xi^i = \Delta s^2$ and are the best candidates for long range perceptual measures.

In the space of Vorobyev and Osorio (1998) the photopic and scotopic cases are treated individually and not shown as a continuum of k/ξ between 1 and 0 - unlike here. Furthermore, in their model the space I have so far described is further elaborated upon. This is the subject of the following section. Before doing this, I will specify the transformation of the quantum catch space to the perceptually uniform space. This is simply a solution to:

$$g_{ij}(\xi)\mathrm{d}\xi^i \mathrm{d}\xi^j = \frac{\delta^l_{ij}\mathrm{d}\xi^i \mathrm{d}\xi^j}{Ad_{(l)}(\xi^l + k_{(l)})} = \delta_{ab}\mathrm{d}\rho^a \mathrm{d}\rho^b \tag{5.4.5}$$

noting that the metric has no cross terms we have for each coordinate:

$$\mathrm{d}\rho^2 = \frac{\mathrm{d}\xi^2}{Ad(\xi + k)}$$

so

$$\frac{\partial\rho}{\partial\xi} = \frac{1}{\sqrt{Ad(\xi + k)}}$$

and (after reintroducing the indices)

$$\rho^i = \int \frac{\mathrm{d}\xi^i}{\sqrt{Ad(\xi^i + k)}} = \sqrt{\frac{4\left(\xi^i + k_{(i)}\right)}{Ad_{(i)}}} \tag{5.4.6}$$

5.5 Projective Models of Colour Vision

The class of projective colour spaces are used to model chromatic-opponent processes - processes where comparisons are made between the excitations of photoreceptors. These, in their most naïve (and common) form are produced by a linear projection of coordinates into a space orthogonal to a 'luminance vector'. This vector is often assumed to be the 'one vector' ($\mathbb{1}$), i.e. the vector whose components in the quantum catch space are all equal to 1. This is fairly trivial in affine spaces, but when the projection is made from a coordinate system where the metric depends on the coordinates, extra considerations must be made.

In an affine space, it suffices to simply consider the null space[12] of the luminance vector. Here, the null space is identical no matter what value the luminance takes. However, as we will see when the metric changes as we move in direction of the luminance vector we have to use a more specific projection, such as the one used by Maxwell (1860), see figures 1.6 and 1.7.

5.5.1 Chromatic Opponency

The most straight forward way of modelling chromatic opponency is by simply considering the vectors spanning the null space of the luminance vector. We simply say that:

$$\rho^i = S^i_j \xi^j \tag{5.5.1}$$

where $[S^i_j]$ is a matrix that spans the null space of the achromatic vector. For example, in human colour vision, it is usually a red-green and a yellow-blue opponency channel, so if:

$$[\xi^i] = \begin{bmatrix} \text{Red} \\ \text{Green} \\ \text{Blue} \end{bmatrix} = \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.5.2}$$

then we have

$$[S^i_j] = \begin{bmatrix} 1 & -1 & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix} \tag{5.5.3}$$

meaning that the opponent mechanisms can be expressed in their usual form (Wyszecki and Stiles, 2000) of:

$$\rho^1 = R - G \qquad \text{and} \qquad \rho^2 = \tfrac{1}{2}(R+G) - B \tag{5.5.4}$$

[12] Let $\mathbf{A}$ be the m-by-n matrix formed by the n-vectors, $\{\mathbf{v}_1 \cdots \mathbf{v}_m\} = V$, then the null space of V (and also of $\mathbf{A}$) is the set of all n-vectors $\{\mathbf{x}\}$ which are solutions to $\mathbf{A}\mathbf{x} = 0$

The general linear transformation of coordinates in equation 5.5.1 could equally be written :

$$\frac{\partial \rho^i}{\partial \xi^j} = S^i_j \tag{5.5.5}$$

There are many possible values for $[S^i_j]$, each one corresponding to a particular set of linear chromatic opponency mechanisms. Here, I consider these mechanisms in general. Applying $[S^i_j]$ gives a new metric tensor $\bar{g}$:

$$\bar{g}_{ab}(\rho) = g_{ij}(\xi)\frac{\partial \rho^i}{\partial \xi^a}\frac{\partial \rho^j}{\partial \xi^b} = g_{ij}S^i_a S^j_b \tag{5.5.6}$$

to get the form used by Vorobyev and Osorio, $\bar{g}$ is expressed as a function of the original coordinates ξ as opposed to the chromaticity coordinates ρ. This has the advantage of the resulting expression being insensitive to the particular choice of $[S^i_j]$ (but not the choice of achromatic vector).

5.5.2 Vorobyev-Osorio Model

Now we are in a position to write an expression for the Vorobyev-Osorio space. It is simply the application of the projection above to the general Poisson space:

$$g_{ij}(\xi) = \frac{\delta^l_{ab} S^a_i S^b_j}{Ad_{(l)}(\xi^l + k_{(l)})} \tag{5.5.7}$$

This works very well in practice, and is theoretically sound when $\xi^i \ll k_{(i)}$ and we can approximate it as affine. However, there is a minor problem in photopic vision. The problem occurs when looking at geodesics, and thus long range colour difference measures. We know that by definition of $[S^j_i]$ as the null space of an achromatic vector $[a^i]$, that:

$$S^j_i a^i = 0 \tag{5.5.8}$$

This means that the geodesic distance d between pairs of points that differ only in the achromatic direction in colour space should always be the same. In other words, there is no cost (in terms of distance between points) for moving either of them in a direction parallel to the achromatic axis. More formally for a metric d, points p and q and scalar constants k and k':

$$d(p,q) = d(p + ka, q) = d(p, q + ka) = d(p + ka, q + ka) \tag{5.5.9}$$

We know that as we move a point away from the origin in the achromatic direction we necessarily increase all of ξ^i and therefore decrease g_{ij}. Restricting the measurement of

distances to within the space perpendicular to a we see that for any $\epsilon > 0$ that we can find a value of k such that:

$$d(p + ka, q + ka) < \epsilon \tag{5.5.10}$$

as the metric g becomes arbitrarily small as we increase k. Thus, as the geodesic is the minimum possible distance under certain constraints, and those constraints do not exclude the method of choosing a path that goes in the direction of achromatic vector to a point arbitrarily far away in terms of coordinates (yet adding no distance), through the null space (adding an arbitrarily small distance) and back in the opposite direction to the achromatic vector to the second point (again adding no distance), then, for any $\epsilon > 0$ we have:

$$d(p, q) < \epsilon \tag{5.5.11}$$

In other words, the geodesic distance between any two points is arbitrary small. In practice this problem is avoided by restricting the choice of colours to an isoluminant plane[13], removing the possibility of points which differ only in the achromatic direction. The model, as it stands, is nonetheless problematic for the methods used in the previous chapter - it is unsuitable for the consideration of long range colour distances.

5.5.3 Honeybee Hexagon Space

We now come to the Honeybee Hexagon space of Chittka (1992): the space I discussed in section 5.3. I must emphasise that Chittka is not the only person to use this kind of methodology, I use his space here firstly because it is a colour space[14] and secondly because he made his assumptions very explicit.

The Honeybee Hexagon Space defines a perceptually uniform coordinate system for the honeybee. The space uses the physiological transform of Backhaus (1992), mapping the quantum catch space to the unit cube (honeybees are trichromats):

$$\rho^i = \left(\frac{\xi^i}{k_{(i)} + \xi^i} \right)^d \tag{5.5.12}$$

where d is taken to be 1. It is then assumed, as in Vorobyev and Osorio (1998) to create a uniform space so that the colour distance Δs is given in a projected space by:

$$\Delta s^2 = \delta_{ij} S^i_a S^j_b \Delta\rho^a \Delta\rho^b \tag{5.5.13}$$

[13] An isoluminant plane is defined with respect to a particular achromatic vector a (usually taken to be $\mathbb{1}$). It is the locus of quantum catches such that $\langle a, \xi \rangle = a \cdot \xi = constant$.

[14] The misconception that it is how a signal that gets transformed, not how a signal and noise gets transformed, is very common in neuroscience.

where $[S_a^i]$ spans the null space of $\mathbb{1}$. This projects the unit cube into the eponymous hexagon.

To get the associated metric we first find the the Jacobian of the coordinate transform:

$$\frac{\partial \rho^a}{\partial \xi^i} = \frac{\delta_i^{am}}{(k_{(m)} + \xi^m)^2} \tag{5.5.14}$$

meaning that the metric in quantum catch space is simply:

$$g_{ij}(\xi) = \delta_{\alpha\beta} S_a^\alpha S_b^\beta \frac{\partial \rho^a}{\partial \xi^i} \frac{\partial \rho^b}{\partial \xi^j} = \frac{S_i^a S_j^b \delta_{ab}^l}{(k_{(l)} + \xi^l)^4} \tag{5.5.15}$$

Whilst it differs from the Osorio-Vorobyev model by some constants and a power, the two are pretty similar. Both of them have a projective part and a part of with the form $1/(A\xi+B)^C$. This is a feature common to all the spaces in this chapter (with the exception of the one to follow, which differs very slightly). This is an important informational feature which I will return to.

However, as I have stated before, we cannot base a colour metric on physiology alone, there must have been some assumptions made. *Post hoc* we can see that this corresponds to Gaussian noise with unit variance centred around the ρ coordinate (up to invariances of the Fisher metric). But, the space was not based on such a theoretical construct, it was simply an application of the transform of Backhaus (1992) with an empirical justification.

However, we can arrive at a very similar result on a (comparatively) rigorous basis. The next section describes how we can get a space that describes photoreceptors that saturate in accordance with the phenomenology of the Bezold-Brüker effect through purely statistical considerations. I will show that the coordinate transform that such an exercise yields is similar, but not equivalent to that used in the Honeybee Hexagon space. The noise based model I will describe is the most detailed in this chapter, so in a sense the following work justifies the less rigorously grounded animal colour space of Chittka (1992) using the principles of more theoretically justified animal colour space of Vorobyev and Osorio (1998).

5.6 A Statistical Saturating Photoreceptor Model

Using the framework I have described it is possible to generate colour space metrics for a saturating photoreceptor from first principles. The space developed by Vos and Walraven (1972a,b) is similar in being based on Poisson processes and exhibiting saturation, it is based on fairly arbitrary assumptions (leading to their exact words being quoted in Wyszecki and Stiles, 2000, p676.). The following is a simplified model of the phototransduction processes in a photoreceptor cell. The simplest model is that described in section

5.4, where the metric is simply determined by shot noise from photons. For a single photoreceptor this yields[15] $g_{ij} = g = \frac{1}{\xi+k}$ where ξ is the quantum catch and k is dark noise arising from spontaneous isomerisations of photopigment molecules.

It is well known that G-protein signalling is a fundamental part of the signal transduction pathway. I will proceed from the general form of signal transduction as described by Hao et al. (2007) (but see also Bao et al., 2010; Heitzler et al., 2009; Linderman, 2009, for a comparison of approaches). The models used by Hao et al. (2007) to model yeast osmosensation have a common structure - each amplification stage can be written as an equilibrium catalysed by the catalytic product of the previous step.

$$\varnothing \rightarrow A \overset{C\downarrow}{\rightleftharpoons} B \rightarrow \varnothing \tag{5.6.1}$$

Which is a general adapting molecular sensor. The equilibrium between A and B is catalysed by C. Adaptation is achieved by the constant creation of A and the concentration proportionate flow of B out of the system. For the purposes of the model, here I will assume there is just a single stage, or that the first stage dominates the statistics of the later stages. Explicitly labelling the rates we have:

$$\varnothing \xrightarrow{k_{\text{creation}}} A \underset{k_{\text{reverse}}}{\overset{k_{\text{forwards}}}{\rightleftharpoons}} B \xrightarrow{k_{\text{decay}}} \varnothing \tag{5.6.2}$$

Before continuing I will make a further approximation - that the system is *adapted*. This can be either expressed as the constraint that[16] $k_{\text{creation}} = k_{\text{decay}}[B]$ where the number of molecules in the system is constant up to Poisson noise or that $k_{\text{creation}}, [B]k_{\text{decay}} \ll k_{\text{reverse}}[B], k_{\text{forwards}}[A]$ and the number of molecules is exactly constant. Both of these assumptions allow us to isolate the catalysed equilibrium, so that we can write it as:

$$A \underset{k_r}{\overset{k_b}{\rightleftharpoons}} B \tag{5.6.3}$$

The two approximations imply that we should treat the relationship between k_r, k_b and k_{creation}, k_{decay}, k_{reverse}, k_{forwards} differently. In the first, k_b should depend on k_{reverse} and k_{forwards}, in the second it shouldn't. Also, in the first approximation we know the expected number of A and B molecules is constant, in the second, we know the exact number is constant. I will take the latter approximation as the basis for this model, noting that we may make it more like the first by including the creation and decay rates in the parameters which I will call λ and μ.

With this in place, this model becomes a type of queueing problem, the statistics of which were developed by Agner Krarup Erlang (a gentle but thorough introduction can

[15] We can drop the indices now we are only talking about one photoreceptor class

[16] I use square brackets to mean 'concentration of'

be found here: Iverson, 2010). We can arrive at a probability distribution for the number of isomerised photopigment molecules in the membrane by considering a process where photopigments molecules are photoisomerised according to a Poisson process with rate λ per active molecule and then recycled with at rate of $\mu = 1/\tau$ where τ is the half-life of the photoisomerised molecule. This assumes the recovery to exponentially distributed, as would be expected from reaction kinetics which are first order in bleached pigment. We begin with a state diagram that looks like figure 5.1.

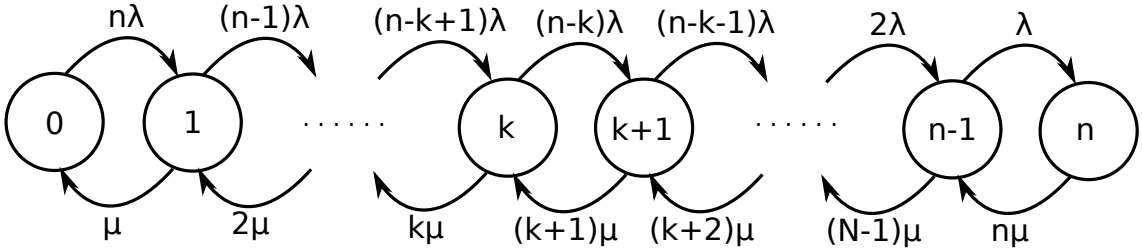

Figure 5.1: State diagram of the number of active photopigment molecules in a photoreceptor membrane.

In this diagram we have states corresponding to each of the possible numbers of active molecules in the membrane - i.e. states numbered as k from 0 to n, where n is the maximum number of isomerised molecules. The number of molecules transitions from a state k to a state $k+1$ proportionally to $n-k$ - the number of active molecules. This is also proportional to the rate of photons hitting the cell: λ. So in total we have a rate of transition for k to $k+1$ of $\lambda(n-k)$. Similarly, we have a recovery process with rate proportional to the number of inactive (bleached) molecules.

From this state diagram we find the probabilities of each state as found in statistical equilibrium. We do this by solving equations for zero net flux between the states, for example, for the state $k = 0$ we have the equation (where p_k is the probability of finding state k):

$$\mu p_1 - n\lambda p_0 = 0 \tag{5.6.4}$$

and for $k = 1$:

$$2\mu p_2 + n\lambda p_0 - (n-1)\lambda p_1 - \mu p_1 = 0 \tag{5.6.5}$$

similarly for all other p_k. We also wish to assure that the probabilities add to one, so:

$$\sum_k p_k = 1 \tag{5.6.6}$$

It is fairly straight-forwards to verify that this system of equations is satisfied by:

$$p_k = \binom{n}{k} \frac{\lambda^k \mu^{n-k}}{(\lambda+\mu)^n} \tag{5.6.7}$$

where $\binom{n}{k}$ is the standard representation of the binomial coefficient. This implies that k is distributed binomially such that:

$$k \sim \textsc{Binomial}\left(n, \frac{\lambda}{\lambda+\mu}\right) \tag{5.6.8}$$

The Fisher metric for this is derived for a general binomial distribution in appendix section D.3. It is given in terms of the coordinate p as:

$$g(p) = \frac{n}{p(1-p)} \tag{5.6.9}$$

so

$$g(\lambda) = \left(\frac{\partial p}{\partial \lambda}\right)^2 g(p) = \left(\frac{\partial p}{\partial \lambda}\right)^2 n\frac{(\lambda+\mu)^2}{\lambda\mu} = \frac{n\mu}{\lambda(\lambda+\mu)^2} \tag{5.6.10}$$

Now we are in a position to introduce more reasonable physical parameters, first we can let $\mu = 1/\tau$ as above, and second we can express λ as the sum of photo-induced isomerisations, ξ, and spontaneous isomerisations ζ, i.e. $\lambda = \xi + \zeta$. So that the metric is now[17]:

$$g(\xi) = g(\lambda - \zeta) = \frac{n\tau}{(\xi+\zeta)(1+\tau(\xi+\zeta))} \tag{5.6.11}$$

and letting $E = \tau(\xi + \zeta)$ we have

$$g(E) = \left(\frac{\partial \xi}{\partial E}\right)^2 g(\xi) = \left(\frac{1}{\tau}\right)^2 \frac{n\tau^2}{E(1+E)^2} = \frac{n}{E(1+E)^2} \tag{5.6.12}$$

In telecommunications theory quantities such as E - a product of a "use rate" and a "recovery rate" is known as the offered traffic - it describes the demand that is put upon a system. In the case of telecommunications this is the rate that calls are made multiplied by how long they last, similarly, here it is how many isomerisations multiplied by how long it takes them to recover. E is a dimensionless characterisation of the system measured in *Erlangs*. It is necessarily positive. A graph of the sensitivity in this coordinate system can be seen in figure 5.2.

This model shows that there is an optimal region for the rate of photopigment replacement for any given isomerisation rate. This is given by $\tau\zeta = 1$ (see figures 5.3 and 5.4).

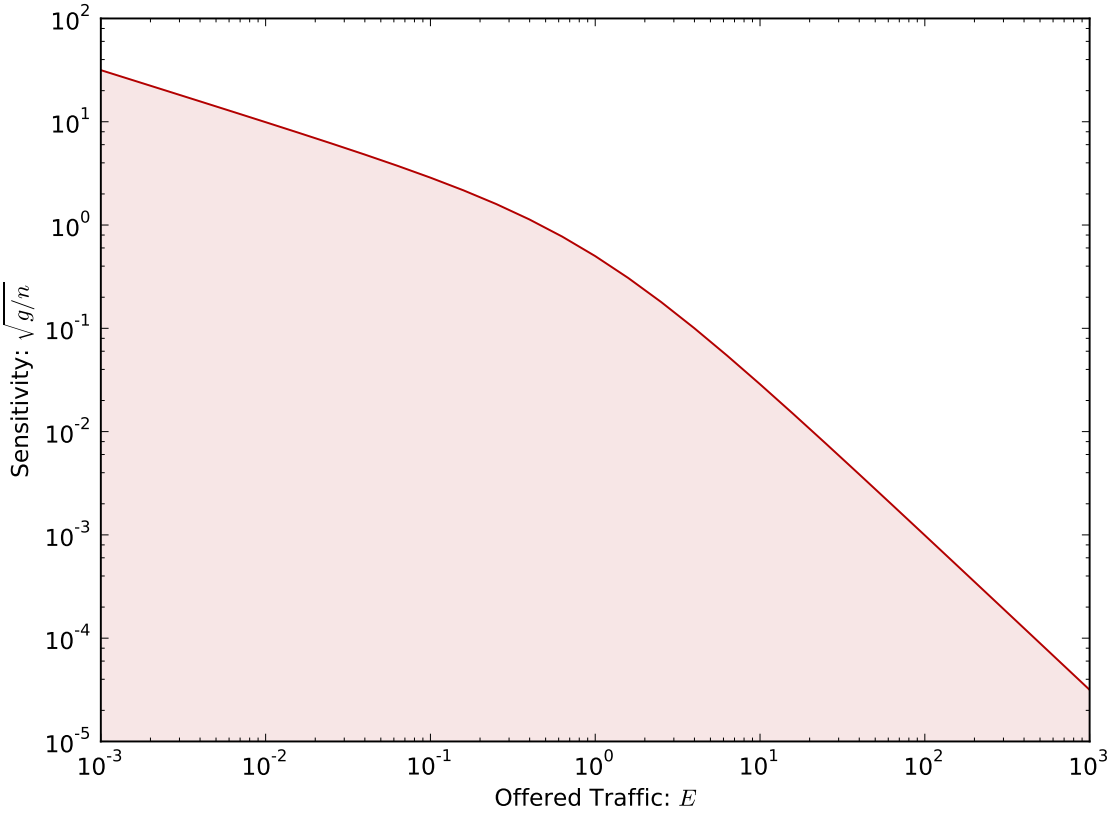

Figure 5.2: The photoreceptor sensitivity with respect to the offered traffic E. In this coordinate system, the sensitivity decreases with increased traffic.

We can go on from here to find the transformation to an affine coordinate system with coordinate ρ. We wish for a metric $g(\rho)$ so that $g(\rho) = 1$:

$$1 = g(\rho) = \left(\frac{\partial E}{\partial \rho}\right)^2 g(E) \tag{5.6.13}$$

so

$$\frac{\partial \rho}{\partial E} = \sqrt{g(E)} = \frac{\sqrt{n}}{(1+E)\sqrt{E}} \tag{5.6.14}$$

and:

$$\rho = \sqrt{n} \int \frac{\mathrm{d}E}{(1+E)\sqrt{E}} \tag{5.6.15}$$

solving the integral by letting $y = \sqrt{E}$ so that $\mathrm{d}E = 2\sqrt{E}\mathrm{d}y$:

$$\tfrac{1}{2} \int \frac{\mathrm{d}E}{(1+E)\sqrt{E}} = \int \frac{\mathrm{d}y}{1+E} = \int \frac{\mathrm{d}y}{1+y^2} \tag{5.6.16}$$

$$= \arctan y = \arctan \sqrt{E} \tag{5.6.17}$$

[17] Here I use the fact that the metric is unchanged on addition of a constant to a coordinate variable: If $y = x + c$ for some constant c then $\frac{\partial x}{\partial y} = 1$ and thus as $g(x) = \left(\frac{\partial y}{\partial x}\right)^2 g(y)$ then $g(y) = g(x + c)$

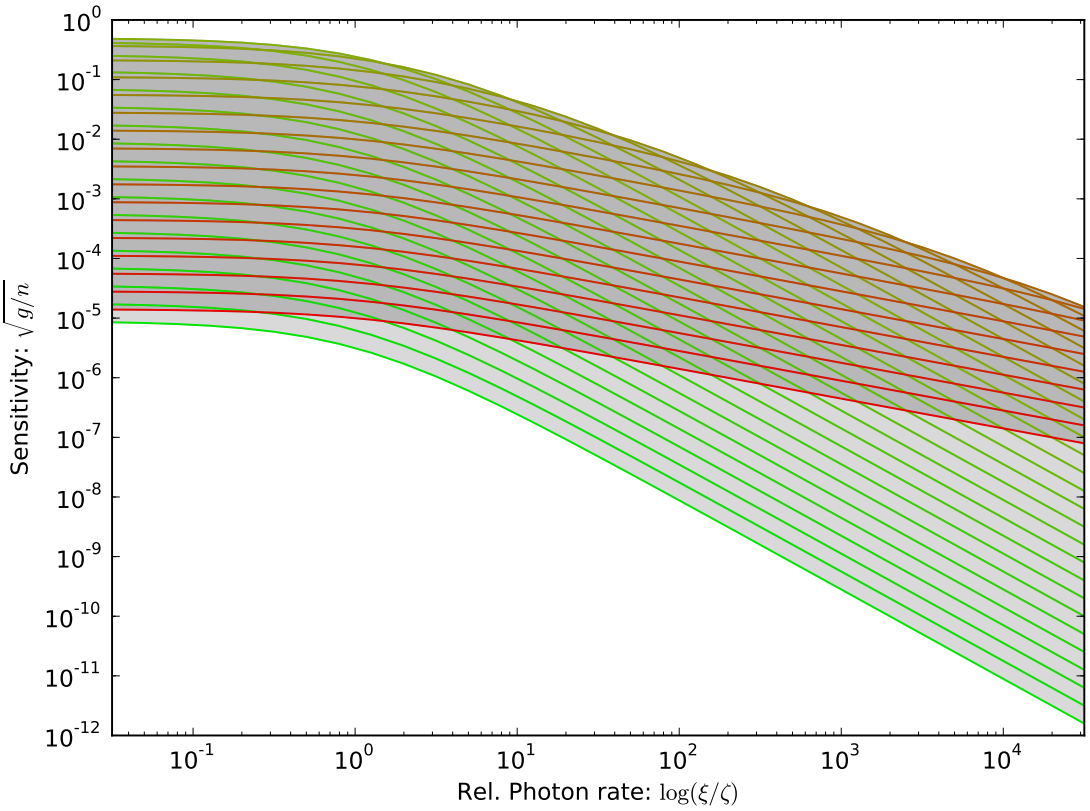

Figure 5.3: Graph of the sensitivity of the saturating photoreceptor model in coordinates normalised by the rate of spontaneous isomerisation. The redness of the lines increases with $1/\tau$: the rate of photopigment recovery. We see at rates faster than where $\tau = 1$ (see figure 5.4) there is a trade off between sensitivity at low photon counts and high photon counts.

So we have an affine coordinate system in ρ as expressed by:

$$\rho = 2\sqrt{n}\arctan\sqrt{E} \tag{5.6.18}$$

or in terms of the physiological parameters:

$$\rho = 2\sqrt{n}\arctan\sqrt{\tau(\xi+\zeta)} \tag{5.6.19}$$

This principled transformation is qualitatively similar to that given by Chittka (1992), but it is not equivalent. It should be expected to behave the same at high luminance values, but in addition it is a noise based metric - grounded in basic physiology.

5.7 General Properties and Classification of Colour Spaces

I will end this chapter with a categorisation of colour spaces by two properties. First by how they behave at high luminosities and second where the point of minimal Fisher

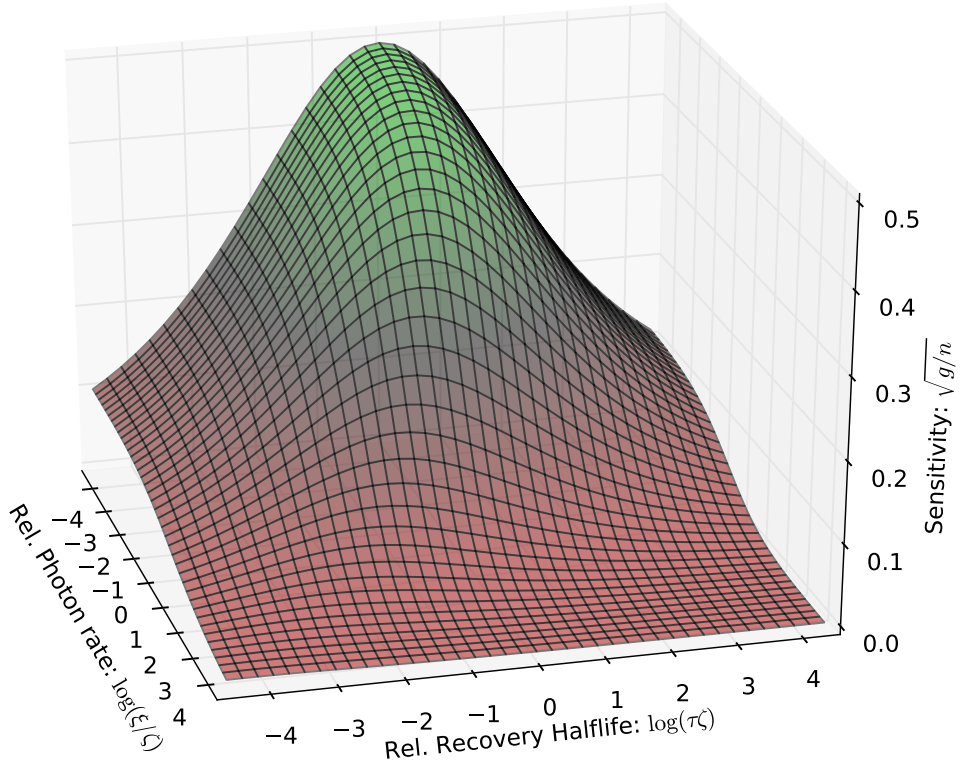

Figure 5.4: Graph of the sensitivity of the saturating photoreceptor model in coordinates normalised by the rate of spontaneous isomerisation. We see there is an optimally sensitive recovery rate where $\tau = 1/\xi$ - i.e. the rate constant of recovery of photopigment is equal to the rate constant for it being destroyed by spontaneous isomerisation. This can be seen in rod and cone photoreceptor cells where the rod cells, with their low spontaneous isomerisation rates, recover from bleaching slowly in comparison with the more noisy cone cells.

information is within any given cross section of the achievable quantum catches.

5.7.1 Saturating or Non-Saturating

It is possible to obtain a very general criterion for whether a colour space describes saturating neurons. Obviously, any colour space that claims to be truly realistic should fulfil these criteria or else imply that our visual pathways have infinite capacity for information transmission.

We can arrive at a fairly natural definition of saturation by considering the mapping of quantum catch coordinates ξ^i to a perceptually uniform coordinate ρ^i. When the intensity of the incident light increases we should expect the information about the light

to decrease due to physiological limitations. The Bezold-Brücker effect implies that the ability to distinguish colour should decrease.

At first it may seem that all we should require is that $\lim_{\xi\to\infty}\frac{\partial\rho}{\partial\xi}=0$, but this is insufficient. Here is why: consider a dichromat with quantum catch coordinates $\xi=(\xi^1,\xi^2)$, choosing two colours α and β we have $\xi_\alpha=(\xi^1_\alpha,\xi^2_\alpha)$ and $\xi_\beta=(\xi^1_\beta,\xi^2_\beta)$. Increasing the intensity of the incident light by a factor of k yields the coordinates $(k\xi^1_\alpha,k\xi^2_\alpha)$ and $(k\xi^1_\beta,k\xi^2_\beta)$ (preserving the ratio of the quantum catches). Now, the euclidean distance between α and β is

$$k\sqrt{(\xi^1_\alpha-\xi^2_\alpha)^2+(\xi^1_\beta-\xi^2_\beta)^2} \tag{5.7.1}$$

so we see that the distance between points in the euclidean metric is multiplicatively increased with intensity. Thus, if we require the distance between these points to tend to zero, not only must the metric decrease with intensity, it must decrease faster than $\log k$ increases, as I will show now.[18]

As we know that $g_{ij}(\rho)=\delta_{ij}$ for a perceptually uniform space, we have:

$$g_{ij}(\xi)\mathrm{d}\xi^i\mathrm{d}\xi^j=g_{ab}(\rho)\mathrm{d}\rho^a\mathrm{d}\rho^b=\delta_{ab}\mathrm{d}\rho^a\mathrm{d}\rho^b \tag{5.7.2}$$

so that for a diagonal metric

$$\rho^a=\int_0^{\xi^a}\sqrt{g_{ii}(\xi)}\mathrm{d}\xi^i+A_{(a)} \tag{5.7.3}$$

for constants A.[19] The criterion for saturation is then:

$$\forall\rho^a:\lim_{\xi^a\to\infty}\rho^a(\xi)\to B_{(a)} \tag{5.7.4}$$

for a different set of constants B.[20] In words, as we increase the brightness towards infinity, the point in a perceptually uniform colour space to which it corresponds should stop moving. We can write this another way using big-O notation[21]:

$$\forall\rho^a:O\left(\rho^a(\xi)\right)\preceq O(1) \tag{5.7.5}$$

[18] Exactly the same argument holds for infinitesimal colour distances (which is in a sense more appropriate).

[19] I have not used the integration constant explicitly until now as it has been of little consequence.

[20] This definition is different to that where we require $\lim_{\xi\to\infty}\frac{\partial\rho}{\partial\xi}=0$ as we cannot assume that the operation $\lim_{\xi\to\infty}\frac{\partial}{\partial\xi}$ is the same as $\frac{\partial}{\partial\xi}\lim_{\xi\to\infty}$. This should be fairly clear as a limit for ξ is never a function of ξ.

[21] In big-O notation $O(f(x))=O(g(x))\iff\lim_{x\to\infty}\frac{f(x)}{g(x)}\neq\pm\infty$ and $\lim_{x\to\infty}\frac{g(x)}{f(x)}\neq\pm\infty$. The relationship signifying symbol '$\preceq$' is '$\leq$' for big-O and should be read as "is lower than or equal to" where $O(x)$ is read "the order of x".

Now, let us take a rather general form for the metric that will cover all those discussed above and more besides. Let $[g_{ij}(\xi)]$ be a diagonal matrix of rational functions r_i of ξ^i so that $g_{ii}(\xi) = r_i(\xi^i) > 0$. A rational function is a quotient formed of two polynomial functions. Here, let us let $P(\xi)$ be an order n polynomial and $Q(\xi)$ be an order m polynomial. If the rational function $r(\xi)$ is the ratio $\frac{P(\xi)}{Q(\xi)}$, then the order of r is $n - m$. This lets us consider each dimension independently, so that we can drop the indices and write equation 5.7.2 as:

$$\frac{P(\xi)}{Q(\xi)}\mathrm{d}\xi^2 = \mathrm{d}\rho^2 \tag{5.7.6}$$

so that the coordinate ρ can be written as:

$$\rho = \int_0^\xi \sqrt{\frac{P(\xi)}{Q(\xi)}}\mathrm{d}\xi + A \tag{5.7.7}$$

Now, it is known that such rational functions are real analytic. It is also true that $\sqrt{x}$ is real analytic when $x \geq 0$. Thus, the composition $\sqrt{\frac{P(\xi)}{Q(\xi)}}$ is also a real analytic function. If a function $f(x)$ is real analytic, it can be represented as a convergent Taylor series and we can consider its integral to be that of a polynomial. It is then safe to say that if:

$$\lim_{x\to\infty} \frac{f(x)}{x^D} \neq 0, \pm\infty \tag{5.7.8}$$

then it is implied that

$$\lim_{x\to\infty} \frac{\int_0^x f(y)\mathrm{d}y}{I(x, D)} \neq 0, \pm\infty \tag{5.7.9}$$

where

$$I(x, D) = \begin{cases} x^{D+1}, & D \neq -1 \\ \log x, & \textit{otherwise} \end{cases} \tag{5.7.10}$$

As the order of $\sqrt{\frac{P(\xi)}{Q(\xi)}}$ is $\frac{n-m}{2}$ then from equation 5.7.8 we have:

$$\lim_{\xi\to\infty} \frac{\sqrt{g(\xi)}}{\xi^{\frac{n-m}{2}}} = C \tag{5.7.11}$$

For $C > 0$ (as g is positive) and $C \neq \infty$ (from 5.7.8). This then implies that from equation 5.7.9

$$\lim_{\xi\to\infty} \frac{\int_0^\xi \sqrt{g(\xi')}\mathrm{d}\xi'}{I\left(\xi, \frac{n-m+2}{2}\right)} = \lim_{\xi\to\infty} \frac{\rho}{I\left(\xi, \frac{n-m+2}{2}\right)} = C \tag{5.7.12}$$

and thus

$$O(\rho(\xi)) = O\left(I\left(\xi, \frac{n-m+2}{2}\right)\right) \tag{5.7.13}$$

This means that the requirement of equation 5.7.4 can be written for a single ρ as:

$$\begin{aligned} O\left(I\left(\xi, \frac{n-m+2}{2}\right)\right) &\preceq O(1) = O\left(\xi^0\right) \prec O(\log \xi) \\ \frac{n-m+2}{2} &< 0 \\ m-n &> 2 \end{aligned} \tag{5.7.14}$$

and when all ρ^as are considered, the assigning of indices to n and m allows equation 5.7.4 to be written in full as either:

$$\forall \rho^a : m_a - n_a > 2 \qquad \text{or} \qquad \min_a \{m_a - n_a\} > 2 \tag{5.7.15}$$

I shall apply this to the spaces discussed so far once I have discussed another property. Then, with that property defined in addition to the one above, I will summarise all[22] the colour spaces I have discussed in this chapter as well as a number of others that I have not yet mentioned.

5.7.2 Relative Information at the Achromatic Point

Now I shall consider a slightly less general class of colour spaces to which most the spaces I have described belong. It is possible to perform the same calculation for the ones that do not fit into this class, but I shall not do so here as this case is sufficiently general for our purposes. Throughout this section I will reintroduce explicit summation in place of the implicit summation of Einstein notation.
This class of models has metrics given by:

$$g_{ij} = \begin{cases} B^i \left(\xi^i + A^i\right)^\mu, & i = j \\ 0, & \textit{otherwise} \end{cases} \tag{5.7.16}$$

where $\{A^i\}$, $\{B^i\}$ and μ are parameters. The parameter μ can be identified with $n-m$ in the previous section.

Here I consider is the infinitesimal volume of Fisher information at points in the colour space. This volume is related to, but not equivalent to, the entropy. It is given at a point ξ by

$$\mathrm{d}V(\xi) = \sqrt{\det\left[g_{ij}(\xi)\right]}\mathrm{d}\xi = v(\xi)\mathrm{d}\xi \tag{5.7.17}$$

Where $\mathrm{d}\xi$ is the volume element at point ξ.[23] Using this we can find the points of maximum and minimum information. The volume factor $v(\xi)$ for this class of models is simple to

[22] though I will ignore projections

[23] The volume element is the outer product of the infinitesimals, i.e. $\mathrm{d}\xi = \mathrm{d}\xi^1 \wedge \mathrm{d}\xi^2 \wedge ... \wedge \mathrm{d}\xi^N$.

calculate as the determinant of a diagonal matrix is the product of its diagonal components:

$$v(\xi) = \prod_k B^k \left(\xi^k + A^k\right)^{\frac{\mu}{2}} \tag{5.7.18}$$

Let us now find the stationary points of the Fisher information volume when constrained to an isoluminant plane. This gives us a way of looking at how information varies with respect to colourfulness and also has the side effect of making the result applicable to the two projective projective models previously mentioned. In this case we take the luminance (achromatic) vector to be the $[a^k]$, yielding a constraint of the form:

$$\sum_k a^k \xi^k = b \tag{5.7.19}$$

and which yields the system of Lagrange multipliers

$$\mathscr{L}(\xi, \lambda) = v(\xi) + \lambda\Lambda(\xi) \tag{5.7.20}$$

$$\Lambda(\xi) = b - \sum_k a^k \xi^k \tag{5.7.21}$$

the stationary points are then given by the constraint ($\frac{\partial \mathscr{L}}{\partial \lambda} = \Lambda(\xi) = 0$) and by:

$$\partial_i \mathscr{L} = \frac{\mu}{2(\xi^i + A^i)} \prod_k B^k \left(\xi^k - A^k\right)^{\frac{\mu}{2}} - \lambda a^i = 0 \tag{5.7.22}$$

so

$$\frac{\mu}{2\lambda} \prod_k B^k \left(\xi^k + A^k\right)^{\frac{\mu}{2}} = a^i(\xi^i + A^i) \tag{5.7.23}$$

As the left hand side is the same for every i, though still not a constant. It is possible to simply replace the whole left hand side with another variable, t:

$$t = a^i(\xi^i + A^i) \tag{5.7.24}$$

which can be written as a vector equation defining a line:

$$\xi = \mathbf{X}t + \mathbf{Y} \tag{5.7.25}$$

with $\mathbf{X} = \left[1/a^i\right]$ and $\mathbf{Y} = \left[-A^i\right]$. Notice that this line - upon which the stationary points lie - only coincides with the luminance vector it is equal to the vector of ones, $\mathbb{1}$ (and A^i are negligibly small). The one vector therefore has a special place in colour theory: Projections based upon the one vector are justified in the sense that the achromatic point corresponds to an information extremum in (hyper)planes spanning the one vectors null space. The previous equation gives:

$$\xi^i = \frac{t}{a^i} - A^i \tag{5.7.26}$$

combining it with the constraint in equation 5.7.19:

$$b = \sum_i \left(t - a^i A^i\right) \qquad \text{so} \qquad t = \frac{b + \sum_i a^i A^i}{\dim \xi} \tag{5.7.27}$$

where $\dim \xi$ is the dimensionality of the parameter space (we are modelling a $\dim \xi$-chromat). This gives the coordinates for the stationary points as:

$$\xi^i = \frac{b + \sum_i a^i A^i}{a^i \dim \xi} - A^i \tag{5.7.28}$$

Let us now look more closely at the stationary point. To determine its nature we calculate the Hessian matrix $[\partial_i \partial_j \mathscr{L}]$. Remembering that $\partial_i v(\xi) = \frac{\mu}{2(\xi^i + A^i)} v(\xi)$, component wise we have:

$$\begin{aligned}
\partial_i \partial_j \mathscr{L} &= \partial_j \left(\frac{\mu}{2(\xi^i + A^i)} v(\xi) \right) - \partial_j \lambda a^i \\
&= \left(\partial_j \frac{\mu}{2(\xi^i + A^i)} \right) v(\xi) + \frac{\mu}{2(\xi^i + A^i)} \partial_j v(\xi) \\
&= \left(-\frac{\mu}{2} \frac{\delta_{ij}}{(\xi^j + A^j)(\xi^i + A^i)} \right) v(\xi) + \frac{\mu}{2(\xi^i + A^i)} \frac{\mu}{2(\xi^j + A^j)} v(\xi) \\
&= \frac{1}{4} \frac{\mu(\mu - 2\delta_{ij})}{(\xi^i + A^i)(\xi^j + A^j)} v(\xi)
\end{aligned} \tag{5.7.29}$$

Reintroducing t for simplicity and evaluating $\partial_i \partial_j \mathscr{L}$ at the stationary point we have:

$$\begin{aligned}
\partial_i \partial_j \mathscr{L} &= \frac{1}{4} \frac{\mu(\mu - 2\delta_{ij})}{(t/a^i)(t/a^j)} \prod_k B^k \left(\frac{t}{a^k} \right)^{\frac{\mu}{2}} \\
&= C a^i a^j \mu(\mu - 2\delta_{ij}) \prod_k (D^k)^{\frac{\mu}{2}}
\end{aligned} \tag{5.7.30}$$

where C and $\{D^i\}$ are positive constants. We can then write:

$$\partial_i \partial_j \mathscr{L} = a^i a^j \mu(\mu - 2\delta_{ij}) f(\mu) \tag{5.7.31}$$

Here $f(\mu) > 0$ which means we can ignore it for the purposes of calculating the nature of the stationary point, only looking at the matrix:

$$M = [a^i a^j \mu(\mu - 2\delta_{ij})] \tag{5.7.32}$$

Also, we only need consider the curvature of the Lagrangian in directions other than the chromatic vector. Projecting this into the null space of the vector $(a^1, a^2 ... a^3)$ (spanned by T) we have:

$$\begin{aligned}
Q &= T^T M T \\
&= [-2\mu(1 + \delta_{ij}) a^i a^j]_{i,j=1...N-1}
\end{aligned} \tag{5.7.33}$$

with the indices covering one less dimension. We can now indirectly look at the eigenvalues. We can show that this fulfils Sylvester's criterion for positive (negative) definiteness, and

therefore all of the eigenvalues are positive (negative). The determinant of the upper m-by-m square matrices is given by:

$$\det\left[-Q_{ij}\right]_{i,j=1\ldots m} = (m-1)2^{m-2}\mu^m \prod_{k=1}^{k=m} (a^k)^2 \tag{5.7.34}$$

Which is always positive when $\mu > 0$, so Q is negative definite in this case, conversely, as changing the sign of μ changes the sign of Q, when $\mu < 0$ the matrix is positive definite.

For simplicity, consider the case when $\mu < 0$ (knowing that the opposite will apply for $\mu > 0$). Negative μ means that the Lagrangian curvature matrix is positive definite in directions within the particular isoluminant plane we are considering. This means that we have a minimum in terms of the infinitesimal volume of Fisher information. This volume is similar to a negative entropy, so minimising it is reducing the information content at that point. The stationary point that is found here is, in a certain sense, the least informative point. Empirically, the necessity for μ to be less than zero can be thought of in terms of terms of discrimination, the sign of μ must be negative if it is to make the correct predictions about spectral sensitivity curves – peaks would become troughs and *vice versa*.

5.7.3 Comparison of Colour Spaces

We now have two qualities that we can use to classify colour space: Whether the achromatic point is an information minimum or an information maximum, and whether they exhibit a Bezold-Brüker type saturation phenomenon. Both of these can be considered on a single axis, $\mu = n - m$. As the model used to investigate information minima is not as general as the one used in the investigation of saturation, to find the position of information minima we must perform some extra calculations for some spaces that do not correspond to the general model. Table 5.1 is a summary of the two global properties. Curvature corresponds to whether the space has a minimum or maximum near the achromatic point - positive when there is a minimum, negative when there is a maximum. Saturating refers to the exhibition of the Bezold-Brüker effect.

Space	μ	**Curvature**	**Saturating**
Gaussian Noise Based Vorobyev and Osorio (1998) (Scotopic) CIE (1931) - XYZ CIE (1976) - $L^*u^*v^*$	0	None	No
Schrödinger (1920) Vorobyev and Osorio (1998) (Photopic)	1	Positive	No
Stiles (Wyszecki and Stiles, 2000) CIE (1976) - $L^*u^*v^*$ - Y coordinate	2	Positive	No
Space described in section 5.6	3	Positive	Yes
Chittka (1992) Chittka (1992) ($d \neq 1$) Vos and Walraven (1972a,b)	4	Positive	Yes

Table 5.1: This table summarises the colour spaces discussed in this chapter. All spaces have either have even information content across the quantum catch coordinates (μ=0) or there is an information minimum at the achromatic point (positive signed embedding curvature in the isoluminant plane). Only a few of the spaces here exhibit saturating effect ($\mu > 2$). For the CIE (1976) entries, the value of μ was assessed by using the order of the coordinate axis in the volume element, it differs between the Y and X, Z coordinates. CIE (1976) is given a value of 0 by equation 5.7.15 but if we consider only the Y value (related to both the green channel and luminance), we have a value of 2.

When considered in quantum catch coordinates all the spaces I have discussed are either flat (Gaussian, Scotopic Vorobyev and Osorio, 1998 and CIE, 1976), or have an information minimum at (or around) the achromatic point. This would seem to be a universal property.

Only a few spaces exhibit a saturating effect (Chittka, 1992; Vos and Walraven, 1972a,b, and mine). These spaces, where $\mu > 2$, are the only ones that would be expected to behave correctly at high luminance values.

5.8 Summary

Using the Fisher metric yields many of the well known colour spaces as specific examples. This correspondence confirms the correctness of its use. During this identification between

the theory of the previous chapter and established colour theory, I have made a number of points that I will reiterate now.

The identification of the metric of colour spaces with the Fisher metric is consonant with the conceptualisation of colour difference as a statistical phenomenon. Indeed, I have shown that when we treat colour as a deterministic phenomenon we obtain a result that neither corresponds to our intuition or to empirical evidence. I have highlighted a case where this has been ignored.

Photoreceptors saturate. A proper model of colour spaces should have perceptually uniform coordinates that tend towards a finite value with respect to the quantum catch coordinates. I have described a model based upon the statistics of molecular interactions which exhibits this behaviour.

All colour spaces have an information minimum in quantum catch coordinates, this is at, or near to, the achromatic point (as I have defined it). It may even be a natural definition of it!

In addition to these points, I would like to add another. When studying animal colour vision to find uniform colour coordinates we should apply the function:

$$\rho(\xi) = k_1 \arctan \sqrt{k_2(\xi + k_3)} \tag{5.8.1}$$

to find perceptually uniform spaces for animals. Here k_1 is the fraction of photoreceptors of a particular type and the constants k_2 and k_3 are physiologically determinable parameters (see above). The number of constants can be reduced by using appropriate approximations, such as $k_3 = 0$ in photic conditions. This would result in space that is is based on solid principles like (Vorobyev and Osorio, 1998) but nonetheless qualitatively similar to that of Chittka (1992). This model is fairly minimal and thus easily applied to all animals.

Part III

Colour in Nature

Chapter 6

The Colourfulness of Signals in Animal Communication

" 'Must a name mean something?' Alice asked doubtfully.
'Of course it must,' Humpty Dumpty said with a short laugh: 'my name means the shape I am – and a good handsome shape it is, too. With a name like yours, you might be any shape, almost.' "

Lewis Carroll
Through the Looking Glass, 1871

In *The Descent of Man, and Selection in Relation to Sex* Darwin talks of organisms being selected for according to judgements of beauty (Darwin, 1871) – an idea which fitted well with his more general investigation into the similarities between man and other animals.

Whilst the notion of æsthetic preference and use of the word 'beauty' in discussions of animal behaviour are now uncommon (excepting Burley and Symanski, 1998; Welsh, 2004), the current study of sexual selection still investigates the same phenomena that concerned Darwin.[1] In modern evolutionary theory Darwin's 'beauty' has become a broad collection of differentiated theoretical constructs (see Endler and Basolo, 1998, for a comprehensive list). But within this pool of theories, the Darwinian principle of female preference for 'attractive' males is still considered to be of wide importance - even if its exact role and meaning is frequently disputed (see e.g. Faivre et al., 2003; Hamilton and Zuk, 1982; Pape

[1] The decline of the word beauty in evolutionary theory most likely stems from neglect of sexual selection arising from the arguments of Wallace (Gayon, 2010) followed by the positivist attitude taken when it was revisited by Fisher (1930).

et al., 1996; Roughgarden et al., 2006; Takahashi et al., 2008; Zahavi et al., 1999, for a broad range of opinions).

Investigating sexual selection as communication necessarily involves the integrated study of psychology, genetics and ecology. Since the emergence of psychology as a science, the consideration of sensory systems has been the most immediately fruitful means of relating the physical domain to the psychological (see e.g. Masin et al., 2009). This too is seen in the investigation of female mate choice and a number of sense-based explanations of evolutionary scenarios have arisen. These include: exploitation of specific or general sensations, such as the use of forced perspective to increase apparent size in bowerbirds (e.g. Endler et al., 2010; Madden, 2003; Schaefer and Ruxton, 2009); exploitation of pre-existing biases, such as signalling using a colour which is innately preferred (e.g. Bravery and Goldizen, 2007; Endler and Day, 2006; Uy, 2004); and the need to form distinct 'channels' of communication in complex sensory environments, such as the differentiation of calls found in Dendrobate frogs (Amézquita et al., 2011; Ryan, 1990).[2]

6.1 Beauty and Judgement

The rejection of the term 'beauty' in the sexual selection literature might at first seem to be because beauty relies on a subject for which a thing is to be beautiful. We can dismiss this explanation pretty quickly, as biology (especially natural history) takes things that should rightly be called subjects as its foundation.

The real difficulty that the notion of beauty has faced, is not the need for a subject, but the need to choose a particular æsthetic theory in which one can judge the æsthetic value of an object to another organism. It seems then, that it is not because anyone thinks that the idea of a subject is silly,[3] but because of the difficulties that we encounter when we ask "What exactly do you mean, Darwin, when you say 'beauty'?" After all, science does not proceed by engaging in philosophy; in its struggle for definitive explanations it replaces difficult concepts such as beauty with more concrete objects of enquiry. Beauty therefore, I hope, has not been thrown away, but has fallen down the cracks between mechanistic explanations.

[2] The formation of preferences for and signals within distinct parts of signal space was preempted in sensory ecology and neuroethology with the concept of Matched Filtering (Franz and Krapp, 2000; Wehner, 1987, 1989) and has been somewhat formalised as a semantic theory by Donaldson-Matasci et al. (2007) (Donaldson-Matasci, 2008, also)).

[3] Although I'm sure some do!

6.1.1 In the Eye of All Beholders

A quality of æsthetic judgements is their normativity. That is, when we find something in good taste we are tacitly or otherwise expecting that others will too ("a claim to validity for all men", Kant 1790, p51). For example, if I tell you that a particular brand of Camembert is the best, I am doing so to help you choose a cheese that is to your liking also. I expect us to share a similar appreciation for cheese. This means that the idea of animals making an æsthetic judgement faces a challenge: that of anthropic bias. Who are we to say what an animal finds beautiful? It seems that talking about a sense of beauty in animals is either very close to, or is, committing the pathetic fallacy.[4]

The Descent of Man does not exist in fear of anthropic bias - it is about demonstrating that man and animals are not so different after all. Indeed, it would not be unfair to say that its business is to claim that certain human norms extend into the animal kingdom. More specifically, when it comes to æsthetic judgements, there are many cases where it is made clear that Darwin uses beauty to mean a judgement that can be shared by both man and animal; for example, he speaks of the good taste of female birds, whom he considers to be more adept in this regard than some humans (Darwin, 1871, p64). I take the stance that Darwin was right in his observations, but I take it upon myself here to give reasons why he should be.

Most generally: if we have an evolutionary advantage in finding pleasure in something in the natural world, why shouldn't the members of species. After all, their lineage has faced the same struggle to exist, they have sensory organs not unlike our own, and they act in the same world that we do. They have the same imperative to learn[5] and they too live in a world too complex to be understood in all its detail. By virtue of just being alive we all must share certain norms.

But the question that we really want an answer for is this: Why would nature conspire to produce something that *I* find appealing?

This is exactly the right question to ask. If we avoid asking it we risk finding ourselves asking why a general principle should apply to everything except ourselves. Norms held by all living things should be held by us, and *vice versa*, we should entertain the possibility that our norms are held by animals, especially when it seems that they act upon them.

This said, the answer to this question has implications beyond that of understanding

[4] Also known as the anthropomorphic, or sentimental fallacy.

[5] Our sense of pleasure and ability to learn are tightly linked. As evidenced at the behavioural level by classical conditioning experiments and at the neural level by studies, in particular, of dopamine neuromodulation (Arias-Carrión et al., 2010; Bressan and Crippa, 2005).

of æsthetic values. The identification and justification of psychological principles that transcends the boundaries between species is of value no matter what label is put on it.

6.2 Signals

The reason that I have taken some time to discuss æsthetic judgement is because there is a step in the discussion of biological signals that relies upon it. When we consider a peacock's train as a signal and claim that its length or the number of eyespots[6] has some kind of significance as a signal we are appealing to the normativity of judgement – we are saying that length or spottiness is something about which peacocks make judgements, and further, that their valuation[7] is in line with our own. We take it as a given that our perception of the world resembles a peahen's in some way; that, for example, their view of the world is not some jumbled up, kaleidoscopic funhouse inhabited by gnomes. One might call it parsimony.[8] But here I discuss it in terms of the normativity of our judgements. When it comes to discussing signals we need to describe things in a way which seems reasonable to us as well as having validity for organism we are describing.

6.2.1 The Handicap Principle

The catch-all justification in biology is adaptation and the Handicap Principle of Zahavi et al. (1999) uses it *as is*. Briefly, it states that sexual signals honestly advertise their bearers fitness by requiring them to pay a cost to produce a signal.

Using this explanation, the peahen has a very direct evolutionary reason to make a particular evaluation - we do not need to know the mechanism by which it does so - only that those peahens whose judgement is calibrated in a direction consonant with natural selection have, in general, a greater progeny. The handicap principle then, like all adaptive explanations, is a direct appeal to natural selection as a norm setter to which living things may benefit from being aligned to. Those who know the ins and outs of natural selection can make a similar judgement; not in the capacity of a peahen, but in the capacity of a biologist who understands what is deemed good and what is deemed bad by nature. One might say that the handicap principle relies on an acquired taste.

But when we first identified the peacock's train as a potential signal we did not use the handicap principle. The handicap principle is effectively agnostic to the mechanism

[6] A 'substantive æsthetic judgement'.

[7] I mean this in the sense of degree, not meaning.

[8] Parsimony is, in fact, completely wrong, as it is a principle that applies to situations which presuppose particular norms.

of the peahen's judgement and we clearly do not share the need to pick a good peacock to mate with. We cannot use it to explain why the train should be something we pick out as a signal in the first place. The best that we can hope to achieve is saying that it is either a massive energetic cost because of its encumbrance, or that it is obvious to predators (invoking normativity with *them*!) thus costly and an honest signal of the peacock's fitness. The latter isn't good enough. This said, judgement that the peacock's tail is conspicuous is clearly very reasonable, but *why is it reasonable*?

6.2.2 Conspicuousness

In many cases it is possible to rank the appearance of organisms from cryptic (camouflaged) to conspicuous (the opposite). Usually this scale is taken to be based upon the ability for a member of some relevant species to be aware of the existence the organism in question (Bradbury and Vehrencamp, 1998). But, we can get a little further towards a satisfactory explanation by considering the ecological psychology of Gibson (1986). To Gibson, objects of perception are defined by their *affordance* - a jug isn't a jug because of its specific geometry, but because it *affords* us the ability to store and pour liquids.

Using this idea we can rephrase the distinction between the cryptic and the conspicuous as a distinction between things that apparently have low affordance and those whose affordance is greater. Judgements of the degree of 'objectness' of something is the same as a judgement about value of a thing in terms of what it can afford, this idea is very similar to conspicuousness.

We are now led to the newly phrased question: "On what basis do we (living things) make judgements of the degree of affordance, and what is it that validates them?"

Like the rest of this thesis, I shall concentrate on the specific case of colour.

6.3 The Purpose of Colour

The distinction between grey and colourful is very intuitive to us, as are the associations we make with it. Grey means uninteresting, colourful means interesting. We all knew what it meant when John Major's puppet in Spitting Image was painted grey, and we naturally expect the festival of Holi to be a lot of fun. For us, this normative judgement is so primal that most of the time we can survive without ever needing to question it. Indeed, it is only when we begin to use it as a judgement that may be shared with other organisms are we forced to enquire into its basis.

Anything that needs to act in a complex world needs to judge what things are important

and what things are not and the thing to note about colour is that when something is colourful (in the sense of J_R in chapter 2) then there is an associated predictive power. This section attempts to justify this claim.

6.3.1 Colourfulness and Sensory Fidelity

At the sensory level colourfulness tells us two things. I have outlined these in chapter 2. Firstly, it tells us, to some extent, how well we can judge the spectrum that produced a given colour. If something is a very strong red, there are relatively few spectra that correspond to it. Secondly, but for similar reasons, it tells roughly how well we can know how it is seen by other observers. Thus colourfulness helps us predict whether other observers will make a similar judgement about a given colour. It is a measure of informativeness, and it is normative to some degree.

What is it that it informs us about? This is the subject of this section, although I have hinted at it in the sections above. So that you may have some idea of where I am going with the proceeding argument, I will simply say: "colourfulness doesn't happen by accident".[9]

Adaptedness

I begin this section with an assertion: If a colour is to be strong there needs to be a force or pressure towards making it so. This pressure can have varying degrees of directness.[10]

The green of a leaf relates directly to its function as an organ which harnesses the energy of photons. The spectrum of chlorophyll is matched to the light which it uses. This is probably as direct an adaptive significance as there is. Nonetheless, there are other cases where physiological role and colour are connected, for example the very direct significance that colour has with respect to the photoprotective role played by melanins (e.g. Kaidbey et al., 1975) and the fluorescent pigments found in coral reefs (Salih et al., 1997, 2000). With less direct significance, there is the colour of various oxygen carriers. The intense colour of blood derives from the sheer mass of oxygen carrier, whose hue is necessitated by the metal complexes that make oxygen binding possible.[11] On the other end of the scale, there is the colouration of tree bark, which seems to have very little direct

[9] Or, at least, it is very unlikely to.

[10] I think I should make it clear at this point that I am not only talking here about the colour of things being an adaptation in itself, but also as the indirect result of some other adaptation.

[11] It seems that the use of metal ions is a universal solution, whether it is haemoglobin, haemocyanin (Halliburton, 1885) or otherwise (Basolo et al., 1975), (with the notable exception of synthetic perflurocarbons Castro and Briceno, 2010).

adaptive significance. With even less significance we have the colour of sedimentary rocks. Then with no relation to biological processes whatsoever we have the colours of igneous rocks, the sky, the moon, and so on. I have deliberately excluded colouration which can be considered as a signal or as camouflage from these examples.

It is in fact rather rare for an organism's colour to not be associated with any adaptation whatsoever. As a corollary: we tend to explain colour in one of three ways, either something is colourful and we see it as a signal or as the indirect result of some extraneous function, or it is not colourful and we see it as camouflage. After all, there are costs and benefits to having any given appearance and every configuration of an organism can be ascribed some utility (Gould and Lewontin, 1979).

At this point we have the beginning of a theory that can explain why it is the strongly coloured animals that are the ones that are signalling. In the examples I have given (barring, possibly, the sky), strong colours are formed when the bearer has some kind of evolutionary function, the stronger the colour the more singular the function. Things that are strongly coloured are because they are affording something.

As it stands the evidence I have presented so far constitutes a fairly weak argument, luckily we can shore it up a little. In colour there are a number of physical constraints that affect our ability to make judgements about what something affords.

Judging what something affords in many ways is the same as judging what something is. We could, if we wanted, use probability theory to make estimates of the affordance of something, just like we could make estimates about any other property. Affordances are, in any case, the properties that define the objects of our perception.

Physical Mixing

When a colour can be a mixture of two colours it is less colourful than the two original colours (see chapter 1) and we can be less certain about what it is too; is it a mixture of these two colours or those two? Should it be thought of as a mixture or as a *singular* colour in itself?

There are many sources of (convex) mixing of colours and it is often used by animals. A remarkable example is the tiger beetle *Cicindela repanda* (Seago et al., 2009) which appears to be a grey that matches the rocks where it lives, but when its skin is viewed with a microscope, we see that the grey is formed by an array of structurally coloured dots, it is an additive mixture. In this case, the grey colour is actually a mixture of other colours. This case of mixing is only one of many (Seago et al., 2009; Welch and Vigneron,

2007, see e.g.). When we mix colours in this way, we expect less colourfulness, to have a strong colour nature must avoid having to mix - in a metaphorical sense, it must go against entropy.

In a similar way to the spatial mixing I have described, a mixture of pigments is generally less colourful. This is known by anyone who has tried to mix paints in an attempt to recover a colour of the original quality. This is exactly the motivation behind the colour system of Alberti mentioned in chapter 1. When biochemical networks do not work to the singular goal of a particular pigment based colour, we should expect less colourfulness.

In appendix E I present a simulation in which in one instance spectra are, in effect, free to mix randomly without any selective pressure. Unsurprisingly, in this simulation neutral drift works towards the species being less colourful.[12]

Back to the case of the tiger beetle. The predator looking at a grey patch of rock does not know if a tiger beetle is there, the converse would not be true were the tiger beetle, for example, bright magenta. The magenta beetle provides certainty. Of course, were it living on magenta rocks, this would not be true. But there is very little physical force or evolutionary pressure in the direction of magenta rocks.

Mixing by Poor Resolution

The geometry of the colour solid described in chapter 1 is a result of poor spectral resolution. When the set of photoreceptors is insufficient to resolve necessary spectral details we see different colours projected onto the same part of the colour space. Whilst it seems that this does not happen very often there is another case related to the additive mixing above - that of poor *spatial* resolution. Simply put, when the eye is incapable of spatially (or temporally) resolving two colours, they are seen as a single colour which is closer to grey by virtue of being an additive mixture. This is why when discussing metamerism in 1 I was careful not to imply that the presence of convex mixtures was about the spectra of objects. Here, we see that the significance comes from the relative probability that some patch of colour is formed from two other distinct colours, this was partly covered in the previous subsection. This is a very general reason why we should find grey things uninteresting. We just cannot resolve their identity. It is often quite remarkable just how much grey there is in the world, histograms of natural images often have far more grey than one would expect, the explanation must be that it is ignored by us before it can even

[12] This appendix also shows how colourfulness can be occur as a result of needing to be generally different in a world with many organisms that are without selective pressure upon their colour.

enter our awareness.

Given what I have said, it easy to see why contrast is useful if one wants to be recognised. The higher the contrast with the background, the lesser the greying effect caused by poor spatial resolution and the more one establishes ones existence as something with unique affordance. For grey organisms, by not having a given affordance with any significant probability, the grey organism asserts that it may as well not exist.

6.3.2 Neutrality

The semiotician and critic Roland Barthes was concerned with exactly the same principle as me, but from a different angle. He took the distinction between colourfulness and greyness as a given that we can all understand so that he could then use it at a metaphor in a number of social commentaries. He was fascinated by the idea of the neutral – a stance that does not take sides. Neutrality, which he associates with a lack of colour,[13] is for him about the lack of signification (Krauss, 2005). Here, I take an interest in the difference between the grey and the colourful for the same reasons.

> And now, I am going to give you the official name of the spilled color, a name printed on the small bottle (as on the others vermilion, turquoise, etc.): it was the color called Neutral [...] Well, I was both punished and disappointed: punished because Neutral spatters and stains (its a type of dull gray-black); disappointed because Neutral is a color like the others, and for sale: [...] the unclassifiable is classified [...]
>
> *Roland Barthes: The Neutral*
>
> *See: Krauss (2005)*

This should clarify my intentions in this chapter, as well as pinpointing a problem that discussions of colourfulness must face. If lack of colourfulness is to be taken as a lack of signification, then we run into difficulties when, upon recognising this, lack of signification becomes a signified at a meta level. We see this phenomenon in discussing camouflage. Things that are not colourful are cryptic in two senses, first, in the passive sense that their rôle is unclear, and second, when crypsis is taken as an adaptation it takes an active sense: being cryptic now signifies an 'unwillingness' to disclose one's rôle. Or, taking the Gibsonian perspective where signified affordances *are* the object - at the first level the cryptic organism does not exist, and at another level it exists as something that makes a claim not to.

[13] It is unfair to say that he associated it with 'grey', which is in fact a colour.

6.3.3 Summary

As a rough summary of this section, on one level we can say that colourfulness signifies a singular affordance, and greyness none in particular, but at a meta level, where existence is asserted, colourfulness is an honest claim to existence and greyness, on the other had, is a dishonest claim to non-existence (by having any given affordance with low probability).

6.4 Some Relevant Cases

There are a number of cases where the theory I have discussed is relevant to discussions within the field of animal colour vision. Here I present two that I feel, in particular, give weight to my stance.

6.4.1 The Colour Vision of Bees

It has been observed that the colour of space honey bees, who are trichromats whose photopigments lie in the UV, Blue and Green regions of the electromagnetic spectrum, is arranged so that the green leaves falls in the same area as grey objects (Chittka, 1992). This is not to say that it is impossible for bees to distinguish green and grey objects, it has been shown that they can (Vorobyev et al., 1999). It is simply that, where green leaves lie on the edge of our colour space, they lie in the centre of the space of bees.

Using the preceding theory it is possible to speculate about why this should be so. If, as I have argued, the uncolourful centre of colour spaces signifies something of little affordance, then it seems right to suggest that there is an evolutionary advantage in putting the colour of leaves into this area, along with everything else which has lesser value to the bee. By making green similar to grey, the bee is free to spend its time investigating more important things like flowers and other bees.

6.4.2 Bowerbirds

Bowerbirds like many corvids display remarkable cognitive abilities, particularly so when it comes to constructing the elaborate structures by which they are named. The male bowerbird produces his bower[14] as part of a mating ritual in which the female enters the bower from one end to see the male displaying at the other. The display takes place atop a collection of stones (or other items) which the male has chosen to enhance his display (Bravery and Goldizen, 2007; Endler et al., 2010; McManus and Weatherby, 1997).

[14] A tunnel made of grass and twigs.

Endler and Day (2006) have performed an experiment testing the colour preference that male bowerbirds have towards the stones they use for constructing their bowers. Whilst they display some preference for greens, the background colour of their environment, their greater preference was towards grey.

The explanation for this, in the terms I have been using, is that they are making a judgement about the stones which deems them to have no clear affordance. They deem the stones they choose uninteresting, be it consciously or otherwise. For the male, who is far from grey, putting themselves in front of such a background is a way of making the females more aware of them. It is well known that such contrast effects are used by bowerbirds in the other modalities, as exemplified by their use of forced perspective (Endler et al., 2010). This is in contrast to other, more well known species of bowerbird, that collect colourful coloured objects, but do not stand over them.

6.5 Conclusion

Colourfulness *is* a pre-existing preference, and *it is* sensory exploitation. It both relates directly to the limitations that the sensory system places upon judgements of affordance as well as taking advantage of the very same faculties that are needed to perform any task in a complex world.

The primary role of the colourfulness of a signal is not to carry any specific meaning, but to assert the bearer's existence as something that should not be overlooked. It is not a standing out from the background by contrasting with it, but a standing out by a claim to being the bearer of utility. Conspicuousness may well be a very good word for this, but I feel it must come with a warning. If taken in the sense I have just described, we may find it very difficult to say what it is in terms of physical properties.

Going back to the idea of normativity. Like the members of any other species we share a need to comprehend the world around us. But our worlds are complex and so few things in them are relevant to us. We have the need to make a judgement about how useful some object is, and we need to do so before we really understand what that thing is. We, and our animal kin alike, must be careful to spend time with only those things that afford us the most. We share this norm.

We also share physiology, in particular we (usually) have eyes. Though there are many designs of eye, all of them are constrained to but a few types of photopigment. We share uncertainty about the meaning of things that are grey, in part because of this common physiology. But also because grey is a colour favoured by processes without a direction -

by “entropic forces”.

So we should expect other organisms to prefer strong colours and we should expect man to find them worthy of comment for precisely the same reason: colourfulness signifies a singular purpose.

So perhaps animals can rightly be said to have a sense of beauty, or at least of taste. Not in the sense of high-minded appreciation for classical music or portraits by old masters, but in the sense that we all share a basic need to give similar values to the world around us. Maybe Darwin’s use of the word beauty is more exacting than it may seem at first glance.

Chapter 7

Structural Colours in Batesian Mimicry

"We live in a rainbow of chaos."

Paul Cézanne

Up to now this thesis has been mostly concerned with brightness and saturation. There is however, a common natural scenario where colour changes with a physical parameter in a way that is better approximated by a hue-like dimension in colour space, rather than brightness or saturation. This is structural colouration.

Structural colouration describes the phenomenon whereby the physical structure of a surface at the microscopic level causes colour through wave interference. There are a large number of specific physical mechanisms, which I will briefly review later. For now though, to get a feel of how it is different from pigments and their properties, consider how the hue of a rainbow changes with the angle relative to the anti-solar point - the hue changes dramatically but the brightness and saturation remain approximately the same. This kind of feature, a change only in hue, makes structural colour noteworthy.

The calculation of structural colours at the physical level is computationally expensive and repeating these computations across ranges of different parameters is impossible in reality. I tackle the complexity of this problem by making a number of simplifying assumptions to produce a model that retains certain important qualitative properties of colour forming structures. I then use this model to investigate what happens in a well known evolutionary scenario - Batesian mimicry. This scenario involves the formation of strong warning colours (aposematic colours) by an indirect pressure from a mimicking species (Franks et al. 2009; Holmgren and Enquist 1999 and implied by Franks and Noble

2004). Batesian mimicry is one of many possible mechanisms by which warning colours can become stronger. Although I focus on Batesian mimicry here, the qualitative difference between the colour forming mechanisms of pigments and structures is no doubt an important part of many evolutionary scenarios.

7.1 Introduction

The study of the evolution of coloured traits is frequently focused on one of two major scenarios: sexual selection - where colour is used in the communication between females and males of the same species; and aposematism - where the colour serves as a warning to predators that eating the colour bearing organism will be costly. This chapter focuses on aposematism (warning colouration) and the related phenomenon of mimicry.

7.1.1 Aposematism

Aposematism is a communication between a prey and predator species. The prey are toxic and they advertise this to the predator. There are numerous examples of aposematism within the specific domain of colour: the strong yellow stripes of the common wasp *Vespa vulgaris* (e.g. Hauglund, 2006), the variously coloured and patterned poison arrow frogs of the *Dendrobatoidea* superfamily (e.g Summers and Clough, 2001), numerous butterflies, notably those in the family *Heliconius* (e.g. Naisbit et al., 2007b) and many *Meloidae* or blister beetles (Nikbakhtzaseh and Tirgari, 2002, e.g.) to name a few well studied examples.

We can say a feature of a prey species is aposematic with varying degrees of confidence. Most minimally, we can assert aposematism when both the prey is toxic and it is reasonable to claim that the predator can identify the feature with which the toxicity is correlated. We can be more confident in our assertion when in addition we can show that the predator's behaviour is aversive and contingent upon the presence of the feature.

7.1.2 Mimicry

Mimicry is often classified into two types: Batesian and Müllerian, which are exploitative and cooperative respectively. However, the classification is a little complex. First of all, mimicry is discussed in terms of a mimic and a model, the model by definition, is less palatable to predators than the mimic. The model is always unpalatable, in the sense that the predators will on average learn to avoid it. This definition of palatability leads quite naturally to the identification of a surface in 'palatability space' to which the

predator is indifferent. In other words, there is a set of attributes of the prey that upon being discovered during predation do not effect the subsequent attack/avoid decision of the predator. This is a natural point to call zero in the palatability scale. The simplest distinction between Müllerian and Batesian mimicry is on this basis: Batesian when the mimic's palatability is less than zero; Müllerian when the palatability is greater than zero.

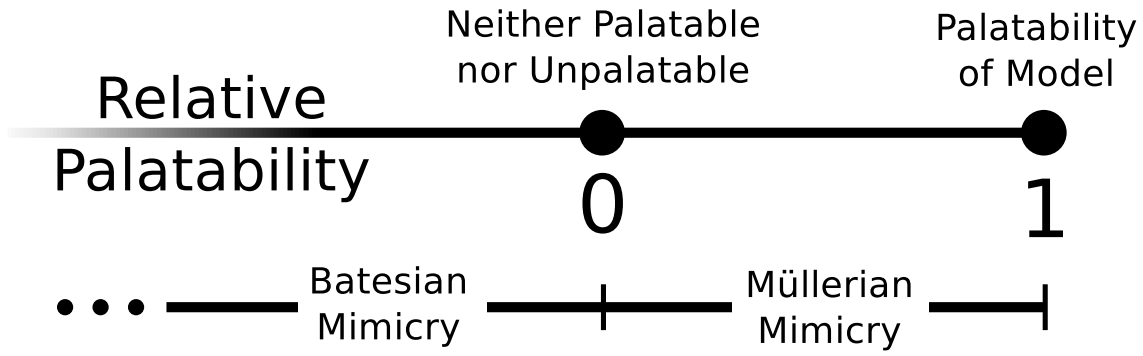

Figure 7.1: Diagram of the palatability scale with Müllerian and Batesian mimicry shown. This is a simplification of figure 7.2

But this simple picture is complicated by a number of factors, such as population size and the details of the learning mechanism of the predator. This leads to the current classification amongst biologists in terms of both palatability and model survival rate (or some similar parameters) (Balogh et al., 2008).

Whilst, the complexities alluded to in figure 7.2 are important for understanding of mimicry. The discussion here assumes *a priori* a Batesian or quasi-Batesian scenario.

7.2 Evolutionary Pursuit in a Colour Solid

An evolutionary chase is exactly what it sounds like: One species is trying to catch up with a second species and the second species is trying to get away from the first.

In the case of Batesian mimicry there is a pressure for the first species (the chaser, the mimic) to resemble the second (the chasee, the model). Likewise, there is an evolutionary pressure for the model to be unlike the mimic. In the case of mimicry, it is assumed that the chase takes place in a perceptual space defined by a third party: the predatory species. Within this feature space, the trajectories of the species is controlled by a multitude of factors including the details of the evolutionary scenario (reproduction, variation), the mapping from genotype to phenotype including environmental factors. All these factors combine to produce a topologically complex network embedded in the predators perceptual

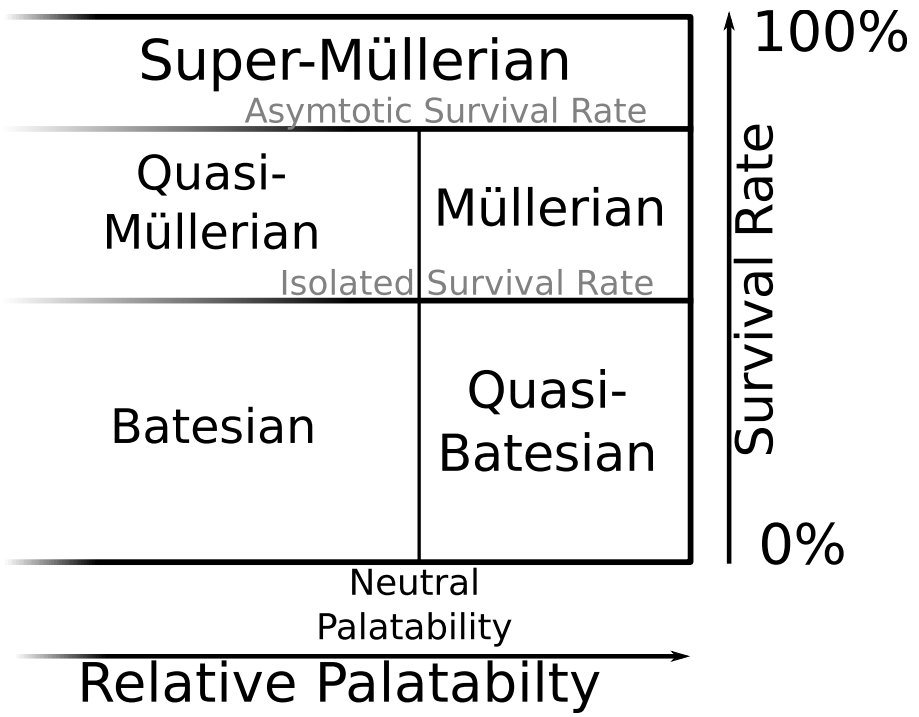

Figure 7.2: A more in depth classification of mimicry (a modification of the diagram in Balogh et al., 2008). Here, both the palatability and the survival rate of the *model* are are used to produce the diagram. The palatability falls into two sections like in figure 7.1. The survival rate is naturally trichotomous: At the bottom, there is a section where the survival rate is less than it would be were the mimic not present. At the top, there is a section where the survival of the model is greater than it would be if the mimic was exactly as palatable as the model. Then there is the area between these two. The super- and quasi- mimicry zones are those that are not covered by the original theories of Müller and Bates, they are however, at least theoretically possible given the appropriate learning mechanisms and population sizes.

space.

When modelling the evolution of coloured signals one may be tempted to assume that some colour space coordinates (RGB for example Holmgren and Enquist, 1999) correspond to continuous traits. This has the effect of removing the complexity of the phenotypic network.[1] The evolutionary landscape becomes featureless, except for the boundary of the colour space.

In the cases where this assumption is made, the boundary of the colour space determines the limiting behaviour of the evolutionary chase. There is a tendency for species to get stuck in the corners. We can explain this with the idea of momentum. The idea is

[1]Instead there is just the standard topology on $\mathbb{R}^n$ generated by the euclidean metric.

simple: the best way for the mimic species to get closer to the model species is by moving directly towards it, the best way for the model species to evade the mimicking species is to move directly away from it. Thus, all things being equal, in a given period of time both species will move the same amount in the same direction. Or more precisely, it is very unlikely that the mimic will move further than its distance to model *plus* the distance the model moved, thus the direction of the vector from mimic to model is preserved. So, with the direction from mimic to model is staying the same during the next time step they both move in the same direction again. Of course, the actual direction and speed of mutation is influenced by many factors. But as a rule of thumb, the evolutionary chase has momentum because once the two species start moving in a given direction in perceptual space they will continue to do so to the degree that physical constraints allow.

7.3 Not Just a Colour Cube

Now that I have discussed the qualitative aspects of how the model and mimic species move though a feature space we can move on to look at the structure of the space itself. To do this I use a model of reflectance spectra as determined by physical parameters. This is intended to highlight the complexities that structural colours might add to the dynamics of mimicry scenarios.

7.3.1 Modelling Reflectance Spectra

There are three components to the model of reflectance spectra. The totality of a reflectance is modelled as a structural reflectance, an absorption due to pigmentation and a potentially imperfect, uniform reflection from a backing layer. This model is only approximate, a full ray based model would include repeated reflections and would still not be as accurate as a full solution to Maxwell's equations. A solution to Maxwell's equation would account for interference effects more realistically, as would be necessary when such small structures are considered. In addition, there would be dependencies on the angle of viewing (iridescence) which will only add complications to the basic point I wish to argue.

In total, the reflectance spectrum $r(\lambda)$ is given by the equation:

$$r(\lambda) = s(\lambda) + b\,(1 - s(\lambda))\,t(\lambda) \tag{7.3.1}$$

Let me break this down a bit.

The first term defining $r(\lambda)$ is a structural reflectance $s(\lambda)$. This reflects light to the observer in a wavelength dependant manner, leaving the transmitted light with a spectrum

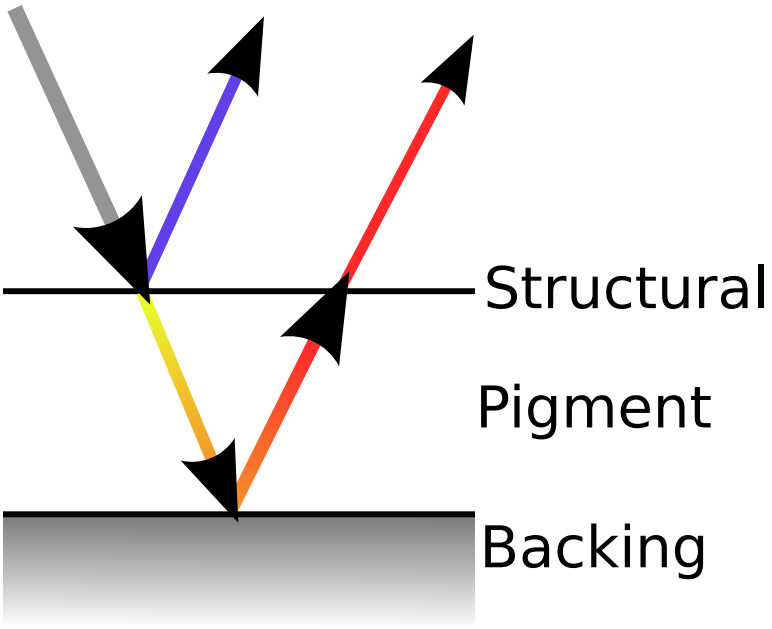

Figure 7.3: Diagram of the physical model used in calculating reflectance spectra. The light incident to the surface is partially reflected by the structural layer. The remaining light then passes through the pigment layer where it absorbed by the medium. It is partially reflected by a backing layer and then passes through the pigment layer again.

given by $(1 - s(\lambda))$. This remaining light passes though an absorbing pigment layer and is partially reflected back (with fraction b), adding to the light reflected by structural layer. At this point in the model we can see that we need the term $bt(\lambda)$ to be different from unity for any of structural colours to make a difference (this reflects the observations of Shawkey and Hill, 2005).

This, as I have indicated above, is a fairly heavy handed simplification of the actual optical phenomenon. For this reason, I will spend some time justifying the simplifications that I have used.

Structural Reflectances

There are a number of features that I have omitted from the model presented here. Structural colours are formed using one of three broad classes of mechanism. Firstly, there are thin films, like in the reflectance from oil on water - a surface can be covered by a thin layer of refractive index different to that of air and the layer beneath - producing an optical cavity that selectively reflects light of different wavelengths. An example of this is the a giant tropical wasp *Megascolia procer javanensis*. It wings appear blue due to a

chitin-melanin[2] backed chitin monolayer of approximately $300nm$ (Sarrazin et al., 2008).

Photonic crystals have a similar mechanism, but their structure does not have the planar symmetry of a film. The three dimensional periodic structure of the material causes a wavelength dependant reflectance - where the period of the structure is a integer multiple of the crystals unit cell size light will be preferably reflected. This is the most common form of structural colouration, it can be see throughout the animal kingdom (see e.g. Welch and Vigneron, 2007, for a brief review).

The bird of paradise Lawes' parotia, (*Parotia lawesii*) uses both a thin film and crystalline photonic structure to produce a startling angle dependant reflectance (Stavenga et al., 2010). Which brings me to the first omission in this model: structural colouration usually appears iridescent - the reflected spectrum is angle dependant. Such effects are impossible to model without providing a the physical relationship between the observer, the observed and the illumination - for this reason I will take it that the spectrum that is modelled is an averaged effect. I also consider the structural layer to be non-absorbing (real refractive indices), all absorbance happens after the initial structural reflection. This single initial reflection is also only an approximation to the real physical scenario.

This means that the structural colours can be removed by taking the maximal wavelength to an extreme on the real line, in effect, becoming a very thick or very thin monolayer. Similarly, it can be thought of as varying the period of the lattice of a photonic crystal (we see this mechanism in peacock feathers for example Zi et al., 2003). This aspect of the model is of vital importance, the inability to modulate the amount of transmittance independently from the reflectance of the structural layer without *radically* changing a physical property is a constraint that necessitates non-linearity in the evolutionary landscape. Although it is not considered here, this would remain true if the degree of reflection was allowed to change as long as it cannot be completely removed and replaced - I assume here that such a case is unlikely as structural colours are intimately related to other physical properties of the surfaces they colour.

I have taken structural reflectances to be given by a Gaussian of unit height centred on μ with constant width (set by σ).

$$s(\lambda) = e^{-\frac{1}{2}\left(\frac{\lambda-\mu_i}{\sigma}\right)^2} \tag{7.3.2}$$

Clearly this lacks formally derived and exacting quantitative realism, but it does have the qualitative features I have mentioned above. This said, they do nonetheless bear some

[2] A structure formed of a composite of chitin and melanin has a broad absorption spectrum as well as differing in (real) refractive index to pure chitin.

resemblance to recorded reflectance spectra - the major difference being the lack of extra peaks. Were they included, these extra peaks only make the evolutionary landscape more complex.

Pigment Absorptions

The general fractional transmittance of an absorbing medium is given by the Beer-Lambert law (Atkins and de Paula, 2001):

$$t(\lambda) = e^{-ca(\lambda)} \tag{7.3.3}$$

where c is proportional to the number of absorbing particles found along the path of a given ray. c is therefor proportional to the thickness of, and the concentration of pigment within, the absorbing layer. $a(\lambda)$ is an absorbance spectrum given in appropriate units to render the term $ca(\lambda)$ dimensionless. Here I take the absorption spectrum $a(\lambda)$ to be representative carotenoid pigments. These pigments are ubiquitous in nature (see for example Faivre et al. 2003; Kodric-Brown 1985 for carotenoids in animals; Armstrong and Hearst 1996 for bacteria; Grotewold 2006 for plants). These pigments generally feature a high absorbance in the short wavelengths and a low absorbance in the long wavelengths. A simple curve with this feature is $1+\tanh(x)$. If we allow it to be 'stretched' and 'moved' relative to wavelength, λ, we can write it in the form:

$$c(\lambda) = \left(1 + e^{-k_g(\lambda - k_c)}\right)^{-1} \tag{7.3.4}$$

where k_g and k_c are parameters that control the gradient at the inflection point and the wavelength of the inflection point respectively. For the purposes here, I keep k_g constant.

Backing Layer

I take the final layer of figure 7.3 to have a spectrally uniform reflectance between zero and one. This corresponds, I argue, to melanin based backing layer (see Mundy, 2005, for a comprehensive review). Though the spectra of melanins are not technically uniform, the approximation as such is not far from the truth for eumelanin - which appears black or deep brown, but it less so for pheomelanin which appears red or orange. However, the backing constant b also represents a reflectivity - in general the amount of light that manages to get reflected back from the internal structure independently of pigment absorption and structural reflectance.

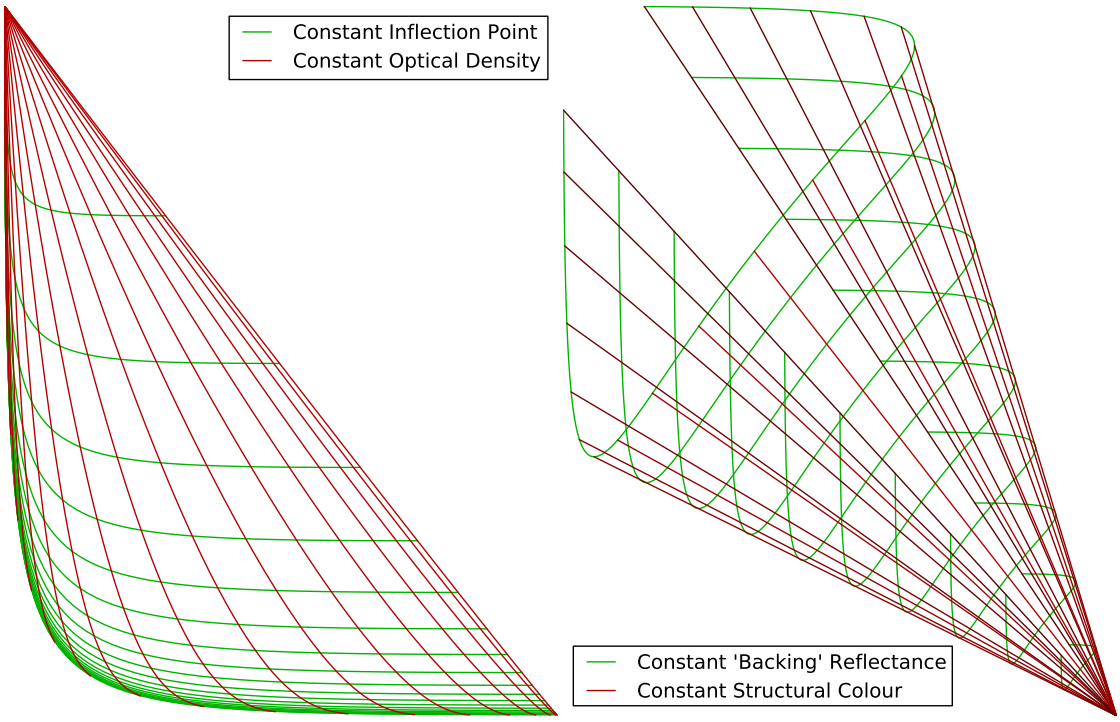

Figure 7.4: A set of colours formed by different physical parameters. The left hand side shows the colours of the pigment over a range of parameters. The right hand shows the structural colour and backing reflectance. The right hand diagram is folded over on itself showing how the structural colours lead to a complex relationship between physical parameters and colour.

Pigment and Backing Layer Features

The pigment and backing layer produce a geometry with nice properties: the space of parameters is projected such that were one to choose a point in colour space, the way one would change the parameters to move in a particular direction in colour space would be well defined. This is not the case with the model stuctural colours. The direction that one moves in colour space given a parameter depends on the value of the parameters to the point where we may need to change a given parameter one way in one case and in the opposite way in another. The parameter for the stuctural colour changes its *orientation* with respect to the colour space at the edges of the cone in the right hand side of figure 7.4.

As I have discussed in chapter 6, the effect of changing pigment concentration is approximately linear. But more importantly, when considered on their own, their parameters keep their orientation with respect to colour.

This means that if we have pigments alone, it is possible to move in a line from any

colour to any other colour - moving locally in the right direction is moving globally in the right direction. The structural colours in this model prevent this property from holding; to get from one colour to another, one may have to begin by going in the opposite direction.

7.4 Continuous Traits

To make any sense of this model as describing part of an evolutionary scenario we are required to make certain assumptions about the relationship between an organisms genes and the colours it displays. Here I assume that the physical parameters of the pigments are continuous traits. We cannot be assured that the physical parameters vary smoothly and mutate linearly in the parameter space, but we could not say this in the case of colour either.

However, with a physical relationship established we *can* say one thing which depends only on the assumption that the change in physical parameters proceeds *via small increments*: We can say that in the scenario above, when given only the information about the local increase in fitness it is impossible for a species to 'know' whether it is going in the right direction to find a global maximum in fitness. This aspect will be essential for the argument that I will make.

7.5 Analysis of the Genotype-Phenotype Mapping

The mapping from genotype to phenotype in this model, whilst a great simplification of the real physical situation, is still complicated and non-linear. The best way of understanding these systems is through visualisation. This allows important features to be observed and explained intuitively. Before I talk about the space, I will explain how I have produced the visualisations so as to provide you with an understanding of their meaning and scope.

7.5.1 Visualising the Genotype-Phenotype Mapping

With the aim of making the features of the geometry clearer I have used a model of a dichromat with Gaussian photoreceptor response functions. The same functions are used for all the visualisations herein. Although different predators will have different types and numbers of visual pigment classes this should not change the gross features of the model.

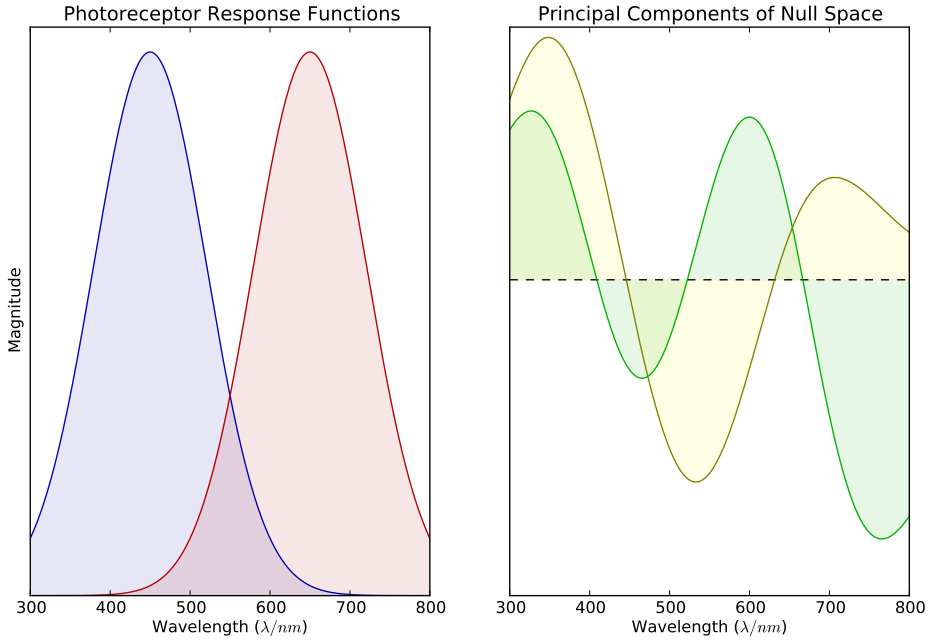

Figure 7.5: Spectra and components used in visualising the evolutionary landscape. The simplified photoreceptor functions are shown on the left. Treating these as vectors, we can find their corresponding null space - i.e. a set of vectors which do not affect the colour - called metameric blacks in colour theory. After uniformly sampling the genetic space and getting the corresponding spectra as $n = 501$ dimensional vectors, projection into the $n - 2$ dimensional null space gives a distribution of the components that leave colour unchanged. The principal components of this space are then used to visualise neutral mutations/metameric changes. To visualise the principal components in the spectrum space the inverse of the transformation into the null space is used, setting the photoreceptor values to zero. See text.

Initial Transformations

The first task is to separate the colour data from the rest of the data present in the spectra. To do this, we take the matrix of the form:

$$\mathbf{M}^T = \left[\begin{array}{ccc|cc} \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{Null}\,\mathbf{W} & & & \mathbf{W} & \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right] \tag{7.5.1}$$

where $\mathbf{W}$ is 2-by-n matrix representing the two photoreceptor spectral sensitivities, Null represents an operator providing a matrix of vectors that span the null space of a matrix. This means that this operation on a spectrum S gives a n-vector whose first $n-2$ compo-

nents do not have any role in determining their corresponding colour (metameric blacks) and whose last two components of the vector are the two colour components, q^1 and q^2. That is to say:

$$\mathbf{MS} = \left[\text{ Metameric Black } \mid \; q^1 \; q^2 \; \right]^T = [\mathbf{B}|\mathbf{Q}]^T \tag{7.5.2}$$

The two colour components were used to make the histogram in figure 7.6 and the two photoreceptor components of 7.8, F.1, F.2 and F.3. To look at the structure of the part of the space which is not involved in determining the colours I have taken the first two principal components of the spectra generated by a small [3] uniform sampling of the parameter space ($\mathbf{P}$). The spectra are transformed according to equation 7.5.2 to yield values in the space with basis vectors given by $\mathbf{B}$. It is in $\mathbf{B}$ that the principal components are calculated. Because of the procedure used here, the principal components of $\mathbf{B}$ are guaranteed to be orthogonal to the photoreceptor vectors in $\mathbf{W}$, providing an efficient representation of the spectrum space. The principal component matrix is then used to reduce $\text{Null}\,\mathbf{W}$ to a 2-dimensional matrix which was used to calculate the values in histograms 7.7 as well as one component of 7.8, F.1, F.2 and F.3. The visualisations herein summarise these principal components p^1, p^2 and quantum catch values q^1 and q^2:

$$\begin{bmatrix} p^1 \\ p^2 \\ q^1 \\ q^2 \end{bmatrix} = \left[\begin{array}{c} (\text{Null}\,\mathbf{W}) \cdot \mathbf{P} \\ \hline \mathbf{W} \end{array}\right] \begin{bmatrix} \vdots \\ \mathbf{S} \\ \vdots \end{bmatrix} \tag{7.5.3}$$

Technicalities of Histogram Production

The two dimensional histograms were calculated using a grid of 640-by-640 bins over the rectangle $[0,1]\times[0,1]$ in the case of the quantum catches (figure 7.6) and $[-6,11]\times[-10,4]$ in the case of the principal components (figure 7.7). These histograms are a result of 2×10^7 simulated spectra. The histogram values are mapped to the luminance channel of an 8-bit image after contrast enhancement using a quadratic weighting ($L \to L^2$).

The three dimensional histograms of 7.8, F.1, F.2 and F.3 use the two colour channels and the first principal component p^1. A similar method on a grid of 100-by-100-by-100 bins

[3] The size of this sample is limited by computer memory. We must keep all the spectra in all their detail to perform a principal component analysis.

spanning the $[0,1] \times [0,1] \times [-6,11]$ rectangle in (q^1, q^2, p^1) was used to make a histogram of 6×10^7 spectra. The histograms were then converted into a surface by placing a point in between the centres of any two histogram cells for which the value of one was zero and the other non-zero. These points were then triangulated using the alpha hull procedure of `MeshLab`[4] with a radius of just greater than $\sqrt{3}$ unit cells. The surface was then simplified to less than 35000 faces using `MeshLab`'s *quadric edge collapse decimation* algorithm and duplicated faces culled (there are many of them).

7.5.2 Visualising the Physiology-Colour Mapping

I have already spoken briefly about the features of the mapping from physiology to colour. But now we have a means of visualising this mapping it is easier to make a few points concerning this space.

Figure 7.6 shows the colour space of the predator as sampled uniformly in the parameter space. Here we see two important features, we see both the triangle on the left hand side of figure 7.4, and the cone on the right. In this histogram it is difficult to resolve the relationship between the two surfaces, but we can see that colours in the bottom left of the image in figure 7.6 can be produced by both pigments and structural colours. Figure 7.7 shows the complexity of the dimensions that have been omitted in figure 7.7.

When we render this in a three dimensional plot, such as 7.8, F.1, F.2 and F.3, we can see two distinct zones (a font zone mostly obscuring a back zone) giving the same colour, but for radically different parameters. The parameter space is embedded in colour space such that there are no possible values for the parameters for areas between the two zones. This means that the colour is not free to change directly from the front zone to the back zone. Figure 7.9 shows two metameric spectra the correspond to each of these zones.

7.6 The Implication for Mimicry and Aposematism

As I discussed earlier, in Batesian mimicry the evolutionary chase between species causes them to have a kind of momentum in colour space (or, more generally, feature space). The consequence of momentum is that both species will, in general, head in a given direction in colour space until they reach the edge of the gamut of possible colours. By this mechanism the aposematic colouration of the model is driven to higher colourfullnesses (see 2). This has been reported in other theoretical studies (Franks et al., 2009; Holmgren and Enquist,

[4] `MeshLab` is a collection of routines for computations involving triangulated surfaces and is "a tool developed with the support of the 3D-CoForm project".

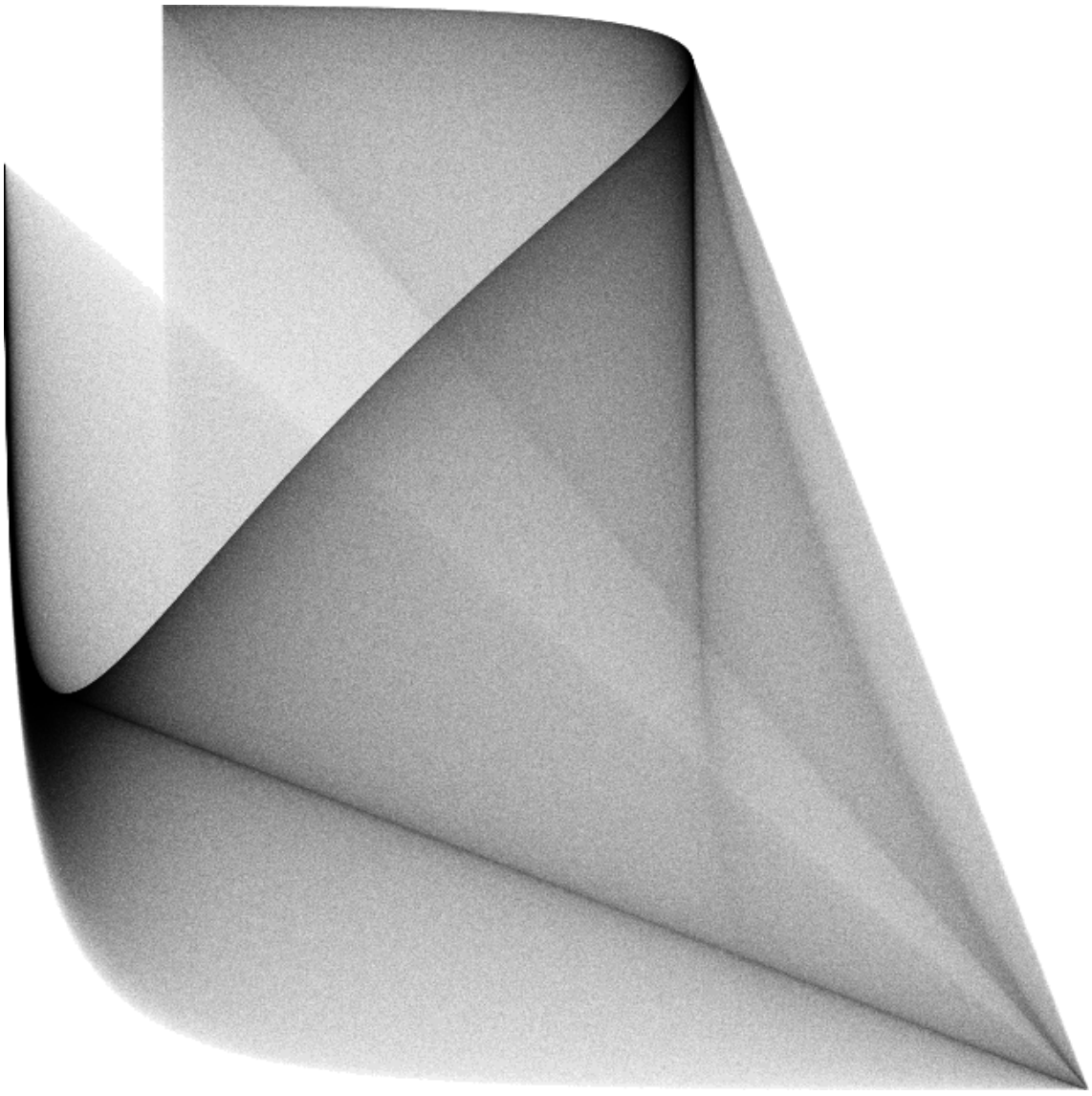

Figure 7.6: A histogram of the quantum catch of the predator's photoreceptors when the genetic space is uniformly sampled. This fits exactly into the unit square $[0, 1] \times [0, 1]$. The values of long wavelength sensitive photoreceptor (red in figure 7.5) span the horizontal axis and short wavelength sensitive (blue in 7.5) the vertical. Here we can see evidence of a complex structure in the genotype-phenotype mapping. It can be seen in figure 7.6 that this corresponds to the folded embedding of the genetic space into it.

1999, e.g.).

However, in the case I present here, the change physiological parameters leading to a particular colour are underdetermined. This means that it is possible for a model species to be similar in appearance to a mimicking species but have rather different underlying physiology (though the possibility for them to be the same is allowed). Indeed, there is

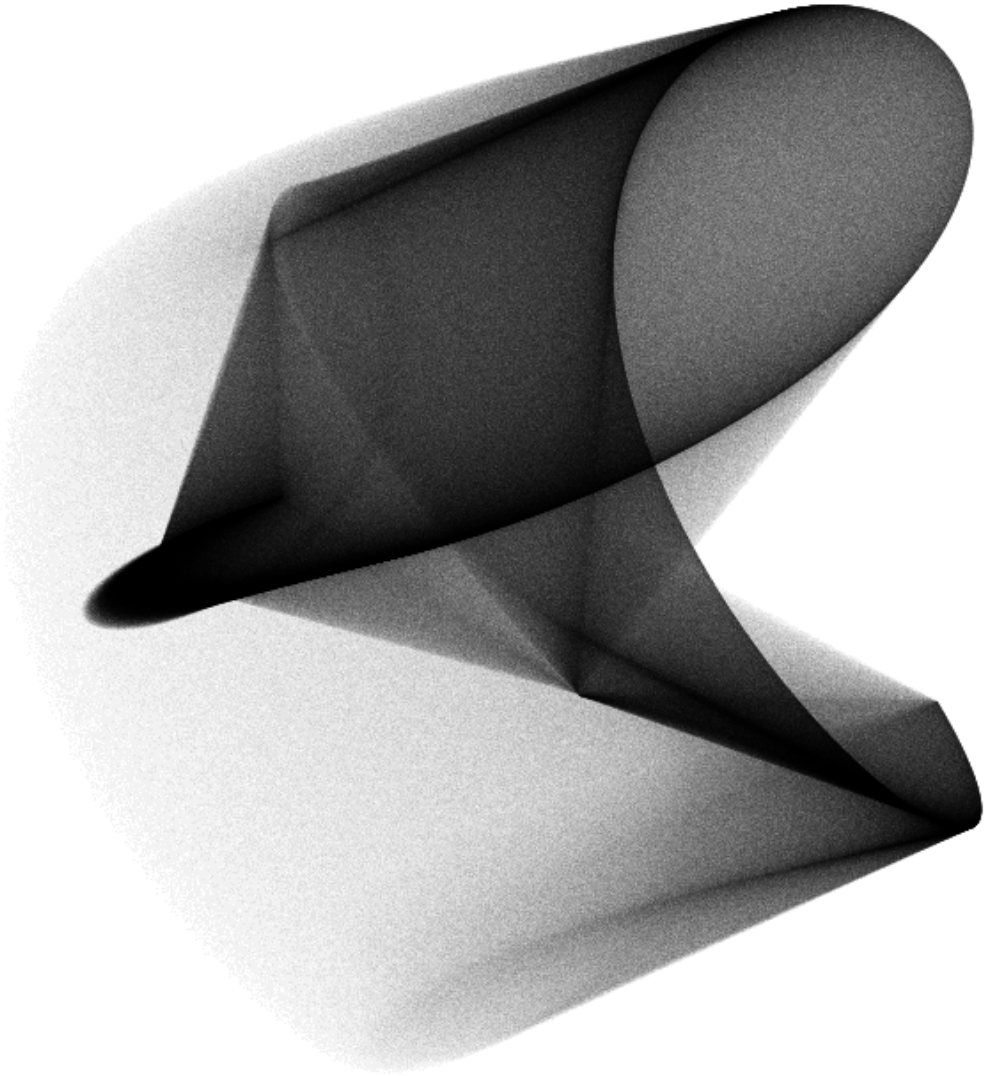

Figure 7.7: The space of neutral mutations - axes are the two principal components described in section 7.5.1. Like before (figure 7.6) we see a complex pattern. Also, we see what appears to be an intersection of two surfaces which forms a concave shape. This concavity will be important later on when we discuss the mechanisms of aposematic colour formation in this model.

neutral drift between all parameters that give the same colour. Because the physiologies may be very different, the change in one physiological parameter that corresponds to a given change in colour may be very different for the two prey species, even if they have the same colour to the predator. Thus, as the model chases the mimic is it very possible that their physiologically diverges instead of converging. As the chase reaches the edge of the colour space the exaggerated difference in physiologies force slightly different extremal colours. The only way of escaping this trap is to make it possible to have large mutations

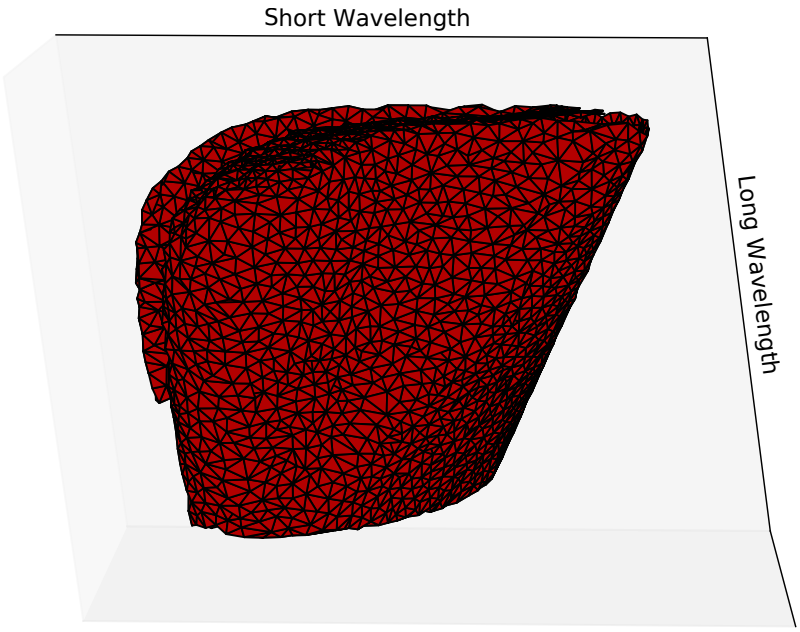

Figure 7.8: Rendering showing the evolutionary trap where one species may be on the front 'leaf' and one on the rear. See appendix F for more renderings.

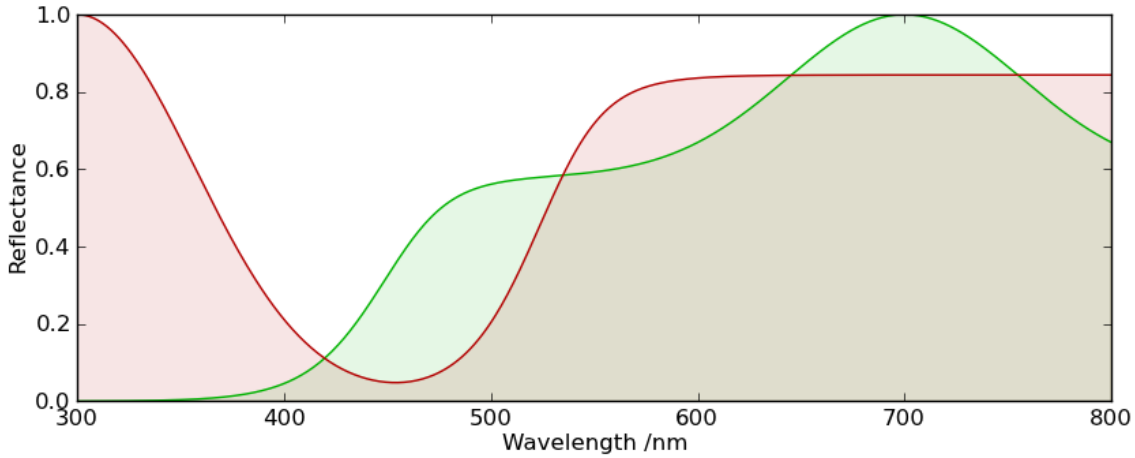

Figure 7.9: Two spectra which have approximately the same colour to the predator. We can see the speak for the structural colour on the left for the red spectrum and on the right for the green spectrum. For the red spectrum to become the same as the green spectrum we would have to move via a desaturated colour and thus, locally, further away from the colour of the green spectrum.

in the physical parameters.

7.7 Discussion

The diversity of physical mechanisms responsible for structural colour is great (Welch and Vigneron, 2007), even before we start combining structural colours (as in Birò et al., 2007; Seago et al., 2009). When this is considered along with the potential complexity of the evolutionary landscapes they create, it is not surprising that where we find structural colouration we find great diversity in colouration. Navigating the landscape of structural colours is necessarily difficult. When it comes to the complexities of the physical basis of colour formation the model presented here is only the tip of the iceberg. The varied visual pigments of different predators, the vast biochemical networks and ecological parameters that affect pigments and high number of possible photonic structures means that there is likely far more potential for finding stable imperfect mimicry.

But as we allow more and more mechanisms, the gamut of colours that can be achieved broadens. With increased mechanistic diversity it becomes easier and easier to produce colours that lie near the middle of the colour solid. The places where models can escape their mimics are pushed further and further to the edges of the colour space.

The difficulty of moving from one part of a colour space to a near by part of the colour space provides a role for "supergenes" (Joron and Mallet, 1998; Mallet, 1989; Mallet and Joron, 1999; Naisbit et al., 2007a) - genetic units that produce a simultaneous change in a large number of physical properties. Although the situation described by these authors is a little different from the one here, the ability to radically switch from one set of physical parameters to others would be expected in a mimic that has evolved to be efficient in modelling other species.

The additive mixing scheme found in tiger beetles (Seago et al., 2009) and butterflies (Birò et al., 2007; Welch and Vigneron, 2007) can be considered as an adaptation to avoid these complexities - an evolution of evolvability (see e.g. Wagner and Altenberg, 1996). By varying the size of mesoscopic zones of structural colours these organisms can linearise the relationship between physiology and colour.

"It never got weird enough for me."

Hunter S. Thomson

Part IV

Bibliography and Appendices

Bibliography

James S. Ackerman. On Early Renaissance Color Theory and Practice. *Memoirs of the American Academy in Rome*, 35(Studies in Italian Art History 1: Studies in Italian Art and Architecture 15th through 18th Centuries):11–44, 1980.

D Alleysson and J Hérault. Photoreceptor nonlinearities can account for the MacAdam ellipses. In *Perception 26 (ECVP Abstract Supplement)*, volume 26, 1997.

Ali Alsam and Graham Finlayson. Matamer sets without spectral calibration. *J. Opt. Soc. Am. A*, 24(9):2505–2512, 2007.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. AMS/Oxford, 2000.

Adolfo Amézquita, Sandra Victoria, Albertina Pimentel, Herbert Gasser, and Walter Hödl. Acoustic interference and recognition space within a complex assemblage of dendrobatid frogs. *PNAS*, 108(41):17058–17063, 2011. doi: 10.1073/pnas.1104773108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1104773108.

Guilio Carlo Argan and Nesca A. Robb. The Architecture of Brunelleschi and the Origins of Perspective Theory in the Fifteenthe Century. *Journal of the Warburg and Courtauld Institutes*, 9:96–121, 1946.

Oscar Arias-Carrión, Maria Stamelou, Eric Murillo-Rodríguez, Manuel Menéndez-González, and Ernst Pöppel. Dopaminergic reward system: a short integrative review. *International archives of medicine*, 3:24, January 2010. ISSN 1755-7682. doi: 10.1186/1755-7682-3-24.

Gregory A Armstrong and John E Hearst. Genetics and molecular biology of carotenoid biosynthesis. *FASEB J.*, 10:228–237, 1996.

Peter Atkins and Julio de Paula. *Physical Chemistry*. Oxford University Press, 7th ed. edition, 2001.

Werner Backhaus. The Bezold-Brücke Effect in the Color Vision System of the Honeybee. *Vision Research*, 32(8):1425–1431, 1992.

Werner Backhaus, Randolf Menzel, and S Kreissl. Multidimensional Scaling of Color Similarity in Bees. *Biological Cybernetics*, C(56):293–304, 1984.

C. V. A. Balogh, G. Gamberale-Stille, and O. Lermar. Learning and the mimicry spectrum : from quasi-Bates to super-Muller. *Animal Behaviour*, 76:1591–1599, 2008. doi: 10.1016/j.anbehav.2008.07.017.

Yifei Bao, Tommy E White, Joseph S Glavy, and Adriana Compagnoni. Application of SPiM to Process Modeling for the Activation Cycle of G-proteins by G-protein-coupled Receptors. Technical report, Stevens Institute of Technology, 2010.

Peter L Bartlett, Michael I Jordan, and Jon D Mcauliffe. Large margin classifiers : convex loss , low noise , and convergence rates. In *Advances in Neural Information Processing Systems 16*, 2004.

Fred Basolo, Brian M. Hoffman, and James A. Ibers. Synthetic Oxygen Carriers of Biological Interest. *Acc. Chem. Res.*, 8(11):384–392, 1975.

J. I. Beare. *Translation of Aristotle's "On Sense and Sensible Objects"*. M.I.T. Internet Archive, 1994.

Mark Bedau. Can Biological Teleology be Naturalized? *The Journal of Philosophy*, 88 (11), 1991.

L.P. Birò, K. Kertész, Z. Vértesy, G.I. Màrk, Zs. Bàlint, V. Lousse, and J.-P. Vigneron. Living photonic crystals: Butterfly scales Nanostructure and optical properties. *Materials Science and Engineering: C*, 27(5-8):941–946, September 2007. ISSN 09284931. doi: 10.1016/j.msec.2006.09.043.

David Bomford. The History of Colour in Art. In Trevor Lamb and Janine Bourriau, editors, *Colour: Art & Science*, pages 7–30. Cambridge University Press, 1995.

Jorge Luis Borges. La biblioteca de Babel. In *El Jardín de senderos que se bifurcan*. 1941.

Jack W. Bradbury and Sandra L. Vehrencamp. *Principles of Animal Communication*. Sinauer Associates, 1998.

Benjamin D Bravery and Anne W Goldizen. Male satin bowerbirds (Ptilonorhynchus violaceus) compensate for sexual signal loss by enhancing multiple display features. *Die*

Naturwissenschaften, 94(6):473–6, June 2007. ISSN 0028-1042. doi: 10.1007/s00114-006-0211-1.

R a Bressan and J a Crippa. The role of dopamine in reward and pleasure behaviour–review of data from preclinical research. *Acta psychiatrica Scandinavica. Supplementum*, 111(427):14–21, January 2005. ISSN 0065-1591. doi: 10.1111/j.1600-0447.2005.00540.x.

Justin Broackes. Substance. *Proceedings of the Aristotelian Society*, pages 131–166, 2006.

G. Buchsbaum and J. L. Goldstein. Optimum Probabilistic Processing in Colour Perception. II. Colour Vision as Template Matching. *Proceedings of the Royal Society B: Biological Sciences*, 205(1159):249–266, August 1979a. ISSN 0962-8452. doi: 10.1098/rspb.1979.0063.

G. Buchsbaum and J. L. Goldstein. Optimum Probabilistic Processing in Colour Perception. I. Colour Discrimination. *Proceedings of the Royal Society B: Biological Sciences*, 205(1159):229–247, August 1979b. ISSN 0962-8452. doi: 10.1098/rspb.1979.0062.

D. Burkhardt and E Maier. The spectral sensitivity of a passerine bird is highest in the UV. *Naturwissenshaften*, 76:82–83, 1989.

N T Burley and R Symanski. "A taste for the beautiful": latent aesthetic mate preferences for white crests in two species of Australian grassfinches. *The American Naturalist*, 152 (6):792–802, December 1998. ISSN 0003-0147. doi: 10.1086/286209.

Camila Irene Castro and Juan Carlos Briceno. Perfluorocarbon-Based Oxygen Carriers: Review of Products and Trials. *Artificial Organs*, 34(8):622–634, 2010.

N. N. Chenstov. *Statistical Decision Rules and Optimal Inference*. AMS, originally edition, 1982.

Lars Chittka. The colour hexagon: a chromaticity diagram based on photoreceptor excitations as a generalized representation of colour oppenency. *J. Comp. Physiol A*, 170: 533–543, 1992.

CIE. ISO 23539:2005(E) / CIE S 010/E:2004 - Photometry - The CIE System of Physical Photometry, 1931.

CIE. ISO 11664-5:2009(E) / CIE S 014-5/E:2009 - CIE 1976 L*u*v* Colour Space and u', v' Uniform Chromaticity Scale Diagram, 1976.

Freeman W. Cope. Derivation of the Weber-Fechner Law and the Loewenstein Equation as the Steady-State Response of an Elovich Solid State Biological System. *Bulletin of Mathematical Biology*, 38:111–118, 1976.

R. T. Cox. Probability, Frequency, and Reasonable Expectation. *Am. Jour. Phys.*, 14:1 – 13, 1946.

R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, 1961.

Charles Darwin. *The Descent of Man and Selection in Relation to Sex*. Amazon, kindle edition, 1871.

Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

Richard Dawkins. *The Extended Phenotype*. Oxford University Press, 1982.

R.R. de Ruyter van Steveninick and S. B. Laughlin. The rate of information transfer a graded-potential synapses. *Nature*, 379:642–645, 1996.

Gustavo Deco and Edmund T Rolls. Decision-making and Weber's law: a neurophysiological model. *The European journal of neuroscience*, 24(3):901–16, August 2006. ISSN 0953-816X. doi: 10.1111/j.1460-9568.2006.04940.x.

Stanislas Dehaene. Symbols and quantities in parietal cortex : elements of a mathematical theory of number representation and manipulation. In *Sensorimotor Foundations of Higher Cognition*, pages 527–574. 2007.

Matina Donaldson-Matasci. *Adaptation in a changing environment: Phenotypic plasticity in response to environmental uncertainty and information*. PhD thesis, 2008.

Matina C Donaldson-Matasci, Michael Lachmann, and Carl T Bergstrom. The evolution of functionally referential meaning in a structured world. *Journal of theoretical biology*, 246(2):225–33, May 2007. ISSN 0022-5193. doi: 10.1016/j.jtbi.2006.12.031.

Michael J Duck. Newton and Goethe on colour : Physical and physiological considerations. *Annals of Science*, 45(5):37–41, 1988.

Charles Lock Eastlake. *Goethe's Theory of Colours*. Frank Cass and co ltd, reprint of edition, 1967.

Albert Einstein. Die Feldgleichungen der Gravitation. In *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin*, pages 844–847.

John A. Endler and Alexandra L Basolo. Sensory ecology, reciever biases and sexual selection. *Trends in ecology & evolution*, 13(10):415–420, 1998.

John A. Endler and Lainy B. Day. Ornament colour selection, visual contrast and the shape of colour preference functions in great bowerbirds, Chlamydera nuchalis. *Animal Behaviour*, 72(6):1405–1416, December 2006. ISSN 00033472. doi: 10.1016/j.anbehav.2006.05.005.

John A Endler, Lorna C Endler, and Natalie R Doerr. Great bowerbirds create theaters with forced perspective when seen by their audience. *Current biology*, 20(18):1679–84, September 2010. ISSN 1879-0445. doi: 10.1016/j.cub.2010.08.033.

Soledad Esteban, Departamento De Química, Facultad De Ciencias, and Senda Rey. Liebig Wöhler Controversy and the Concept of Isomerism. *Journal of Chemical Education*, 85 (9):1201–1203, 2008.

Bruno Faivre, Arnaud Grégoire, Marina Préault, Frank Cézilly, and Gabriele Sorci. Immune activation rapidly mirrored in a secondary sexual trait. *Science (New York, N.Y.)*, 300(5616):103, April 2003. ISSN 1095-9203. doi: 10.1126/science.1082026.

A. A. Fedotov, P. Harremoes, and F. Topsœ. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, June 2003. ISSN 0018-9448. doi: 10.1109/TIT.2003.811927.

Gaoyang Feng and David H Foster. Predicting frequency of metamerism in natural scenes by entropy of colors. *Journal of the Optical Society of America*, 29(2):A200–A208, 2012.

Graham D Finlayson and Peter Morovic. Metamer sets. *J. Opt. Soc. Am. A*, 22(5): 810–819, 2004.

Ronald Almyer Fisher. *The Genetical Theory of Natural Selection.* Public Domain Resource, 1930.

D W Franks, G D Ruxton, and T N Sherratt. Warning signals evolve to disengage Batesian mimics. *Evolution*, 63(1):256–267, 2009.

Daniel W Franks and Jason Noble. Batesian mimics influence mimicry ring evolution. *Proceedings. Biological sciences / The Royal Society*, 271(1535):191–6, January 2004. ISSN 0962-8452. doi: 10.1098/rspb.2003.2582.

M O Franz and H G Krapp. Wide-field, motion-sensitive neurons and matched filters for optic flow fields. *Biological cybernetics*, 83(3):185–97, September 2000. ISSN 0340-1200.

A S French, M J Korenberg, M J arvilehto, E Kouvalainen, M Juusola, and M Weckstr om. The dynamic nonlinear behavior of fly photoreceptors evoked by a wide range of light intensities. *Biophysical journal*, 65(2):832–9, August 1993. ISSN 0006-3495. doi: 10.1016/S0006-3495(93)81116-0.

Jean Gayon. Sexual selection: Another Darwinian process. *Comptes rendus biologies*, 333 (2):134–44, February 2010. ISSN 1768-3238. doi: 10.1016/j.crvi.2009.12.001.

James J. Gibson. *The Ecological Approach To Visual Perception*. Psychology Press, new editio edition, 1986.

Steven J. Gould and R. C. Lewontin. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptionist programme. *Proc R. Soc. Lond. B*, 205:581–598, 1979.

Erich Grotewold. The genetics and biochemistry of floral pigments. *Annual review of plant biology*, 57:761–80, January 2006. ISSN 1543-5008. doi: 10.1146/annurev.arplant.57.032905.105248.

W. D. Halliburton. On the Blood of Decapod Crustacea. *J. Physiol.*, 6(6):300–335, 1885.

William D. Hamilton and Marlene Zuk. Heritable True Fitness and Bright Birds: A Role for Parasites? *Science*, 218:385, 1982.

N Hao, M Behar, T C Elston, and H G Dohlman. Systems biology analysis of G protein and MAP kinase signaling in yeast. *Oncogene*, 26(22):3254–66, May 2007. ISSN 0950-9232. doi: 10.1038/sj.onc.1210416.

R. P. Hardie and R. K. Gaye. *Translation of Aristotle's "Physics"*. M.I.T. Internet Archive, 1994.

K. Hauglund. Responses of domestic chicks (Gallus gallus domesticus) to multimodal aposematic signals. *Behavioral Ecology*, 17(3):392–398, January 2006. ISSN 1045-2249. doi: 10.1093/beheco/arj038.

Domitille Heitzler, Pascale Crépieux, Anne Poupon, Frédérique Clément, François Fages, and Eric Reiter. Towards a systems biology approach of-G protein-coupled receptor signalling: challenges and expectations. *Comptes rendus biologies*, 332(11):947–57, November 2009. ISSN 1768-3238. doi: 10.1016/j.crvi.2009.09.002.

Noél M A Holmgren and Magnus Enquist. Dynamics of mimicry evolution. *Biological Journal of the Linnean Society*, 66:145–158, 1999.

Edmund Husserl. *Philosophy of Arithmetic: Psychological and Logical Investigations*. Dallas wil edition, 1891.

Villy B. Iverson. *Teletrafic Engineering and Network Planning*. Technical University of Denmark (Book Accompanying Lecture Course), 2010.

E. T. Jaynes. *Probability Theory: The Logic of Science*. 1994.

M Joron and J L Mallet. Diversity in mimicry: paradox or paradigm? *Trends in ecology & evolution*, 13(11):461–6, November 1998. ISSN 0169-5347.

V. R. R. Jose, R. F. Nau, and R. L. Winkler. Scoring Rules, Generalized Entropy, and Utility Maximization. *Operations Research*, 56(5):1146–1157, September 2008. ISSN 0030-364X. doi: 10.1287/opre.1070.0498.

Benjamin Jowett. *Translation of Plato's "Theaetetus"*. M.I.T. Internet Archive, 1871.

K.H. Kaidbey, P.P. Agin, R.M. Sayre, and A.M. Kligman. Photoprotection by melanin – a comparison of black and Caucasian skin. *J. Am. Acad. Dermatol*, 1(3):249–260, 1975.

Immanuel Kant. *Critique of Judgement*. Hackett Publishing Co, Inc, 1790.

Astrid Kodric-Brown. Female preference and sexual selection for male coloration in the guppy (Poecilia reticulata). *Behavioral Ecology and Sociobiology*, 17:199–205, 1985.

A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York, 1956.

Hisaharu Koshitaka, Michiyo Kinoshita, Misha Vorobyev, and Kentaro Arikawa. Tetrachromacy in a butterfly that has eight varieties of spectral receptors. *Proceedings of the Royal Society B: Biological Sciences.*, 275(1637):947–54, April 2008. ISSN 0962-8452. doi: 10.1098/rspb.2007.1614.

Rosalind Krauss. From The Neutral: Session of March 11 , 1978. *October*, 8(112):3–22, 2005.

S . Kullback and R . A . Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Solomon Kullback. A Lower Bound for Discrimination Information in Terms of Variation. *IEEE Transactions on Information Theory*, page 127, 1966.

Johann Heinrich Lambert. Beschreibung einer mit Calauischem Wachse ausgemalten Farben-Pyramide, 1772.

Guy Lebanon. Axiomatic Geometry of Conditional Models. *IEEE Transactions on Information Theory*, 51(4):1283–1294, 2005.

C. I. Lewis. The Given Element in Empirical Knowledge. *Philosophical Review*, 61(2): 168–175, 1952.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. ISSN 00189448. doi: 10.1109/18.61115.

Jennifer J Linderman. Modeling of G-protein-coupled receptor signaling pathways. *The Journal of biological chemistry*, 284(9):5427–31, February 2009. ISSN 0021-9258. doi: 10.1074/jbc.R800028200.

Alexander D. Logvinenko. An object-colour space. *Journal of vision*, 9(11):Article 5, 2009.

Gábor Lugosi and Nicolas Vayatis. On the Bayes Consistency of Regular Boosting Methods. *Annals of Statistics*, 32(1):30–55, 2004.

R. Luther. Aus dem Gebiet der Farbreizmetric. *Ztschr. Techn. Phys.*, 8:997 – 1012, 1927.

J Madden. Preferences for coloured bower decorations can be explained in a nonsexual context. *Animal Behaviour*, 65(6):1077–1083, June 2003. ISSN 00033472. doi: 10.1006/anbe.2003.2126.

B Y J Mallet. The genetics of warning colour in Peruvian hybrid zones of Heliconius erato and H . melpomene. *Genetics*, L(236):163–185, 1989.

James Mallet and Mathieu Joron. Evolution of Diversity in Warning Colour and Mimicry: Polymorphisms, Shifting Balance, and Speciation. *Annu. Rev. Ecol. Syst.*, 200X(30): 201 – 233, 1999.

Benoit B. Mandelbrot. *The Fractal Geometry of Nature*. W.H.Freeman & Co Ltd, 1982.

Sergio Cesare Masin, Verina Zudini, and Mauro Antonelli. Early alternative derivations of fechners law. *Journal of the History of the Behavioral Sciences*, 45(1):56–65, 2009. doi: 10.1002/jhbs.

James Clerk Maxwell. Theory of Compound Colours, and the Relation of Colours of Spectrum. *Philosophical Transactions of the Royal Society (London)*, 150:57–84, 1860.

C. McManus and I. P. Weatherby. the Golden Section and the Aesthetics of Form and Composition: a Cognitive Model. *Empirical Studies of the Arts*, 15(2):1–1, July 1997. ISSN 0276-2374. doi: 10.2190/WWCR-VWHV-2Y2W-91EE.

Maurice Merleau-Ponty. *Phenomenology of Perception*. Routledge, 1945.

Nicholas I Mundy. A window on the genetics of evolution: MC1R and plumage colouration in birds. *Proceedings of The Royal Society B.*, 272(1573):1633–40, August 2005. ISSN 0962-8452. doi: 10.1098/rspb.2005.3107.

Russell E Naisbit, Chris D Jiggins, and James Mallet. Mimicry: developmental genes that contribute to speciation. *Evolution & development*, 5(3):269–80, 2007a. ISSN 1520-541X.

Russell E Naisbit, Chris D Jiggins, and James Mallet. Mimicry: developmental genes that contribute to speciation. *Evolution & development*, 5(3):269–80, 2007b. ISSN 1520-541X.

Sérgio M. C. Nascimento, Flávio P. Ferreira, and David H Foster. Statistics of spatial cone-excitation ratios in natural scenes. *Journal of the Optical Society of America*, 19 (8):1484–1490, 2002.

Issac Newton. Of Colours, 1665.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f -divergences. *The Annals of Statistics*, 37(2):876–904, April 2009. ISSN 0090-5364. doi: 10.1214/08-AOS595.

M. R. Nikbakhtzaseh and S Tirgari. Blister Beetles (Coleoptera: Meloidae) in Nahavand County (Hamedan Province , Iran) and Their Ecological Relationship to Other Coleopteran Families. *Iranian J. Publ. Health*, 31(1-2):55–62, 2002.

N Nyberg. Zum Aufbau des Farbenkoerpes im Raume aller Lichtempfindumgen. *Ztschr. Phys.*, 28:406–419, 1928.

OED. Online Historical and Etymological Dictionary, 2012.

N Ohta and G Wyszecki. Theoretical chromaticity-mismatch limits of metamers viewed under different illuminants. *Journal of the Optical Society of America*, 65(3):327 – 333, 1975.

J K O'Regan and A Noë. A sensorimotor account of vision and visual consciousness. *The Behavioral and brain sciences*, 24(5):939–73; discussion 973–1031, October 2001. ISSN 0140-525X.

Anders Pape, Marion Petrie, and Marie Curie. Condition dependence, multiple sexual signals, and immunocompetence in peacocks. *Behavioral Ecology*, 13(2):248–253, 1996.

C. S. Peirce. How to Make Our Ideas Clear. *Popular Science Monthly*, 12:286–302, 1878.

Heather D. Penney, Christopher Hassall, Jeffrey H. Skevington, Kevin R. Abbott, and Thomas N. Sherratt. A comparative analysis of the evolution of imperfect mimicry. *Nature*, 483(7390):461–464, March 2012. ISSN 0028-0836. doi: 10.1038/nature10961.

David L Philipona and J Kevin O'Regan. Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties. *Visual Neuroscience*, 23:331–339, 2006.

M.M. Postnikov. *Encyclopaedia of Mathematical Sciences: Geometry VI: Riemannian Geometry*. Faktorial, Moscow, 1998.

David C Queller and Joan E Strassmann. Beyond society: the evolution of organismality. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1533):3143–55, November 2009. ISSN 1471-2970. doi: 10.1098/rstb.2009.0095.

Howard Rachlin. Teleological Behaviorism. *American Psychologist*, 47(Ii):1371–1382, 1992.

Hayne W Reese. Teleology and Teleonomy in Behavior Analysis. *The Behavior Analyst*, 17:75–91, 1994.

Alfréd Rényi. On Measures of Entropy and Information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1960.

Joan Roughgarden, Meeko Oishi, and Erol Akçay. Reproductive social behavior: cooperative games to replace sexual selection. *Science (New York, N.Y.)*, 311(5763):965–9, February 2006. ISSN 1095-9203. doi: 10.1126/science.1110105.

Michael J. Ryan. The Sensory Bias of Sexual Selection for Complex Calls in the Túngara Frog, Physalaemus Pustulosus (Sexual Selection for Sensory Exploitation). *Evolution*, 44(2):305–314, 1990.

Anya Salih, Ove Hoegh-guldberg, and Guy Cox. Photoprotection of Symbiotic Dinoflagellates by Fluorescent Pigments in Reef Corals. In *Proceedings of the Australian Coral Reef Society 75th Anniversary Conference*, number October 1997, pages 217–230, 1997.

Anya Salih, Anthony Larkum, Guy Cox, Michael Kühl, and Ove Hoegh-Guldberg. Fluorescent pigments in corals are photoprotective. *Nature*, 408:850–853, 2000.

A. C. Sanderson, W. M. Kozak, and T. W. Calvert. Distribution Coding in the Visual Pathway. *Biophysical Journal*, 13:218–244, 1973.

Michaël Sarrazin, Jean Vigneron, Victoria Welch, and Marie Rassart. Nanomorphology of the blue iridescent wings of a giant tropical wasp Megascolia procer javanensis (Hymenoptera). *Physical Review E*, 78(5), November 2008. ISSN 1539-3755. doi: 10.1103/PhysRevE.78.051902.

H Martin Schaefer and Graeme D Ruxton. Deception in plants: mimicry or perceptual exploitation? *Trends in ecology & evolution*, 24(12):676–85, December 2009. ISSN 0169-5347. doi: 10.1016/j.tree.2009.06.006.

Erwin Schrödinger. Grundlinien einer Theorie der Farbenmetrik im Tagessehen. *Ann. Physik*, 63:408, 1920.

Ainsley E Seago, Parrish Brady, Jean-Pol Vigneron, and Tom D Schultz. Gold bugs and beyond: a review of iridescence and structural colour mechanisms in beetles (Coleoptera). *Journal of the Royal Society, Interface / the Royal Society*, 6 Suppl 2(October 2008): S165–84, April 2009. ISSN 1742-5689. doi: 10.1098/rsif.2008.0354.focus.

C E Shannon. A Mathematical Theory of Communication. *Mobile Computing and Communications Review (reprint)*, 5(I):3–55, 1948.

Alan E. Shapiro. *Fits, Passions and Paroxysms. Physics, method and chemistry and Newton's theories of colored bodies and fits of easy reflection.* Cambridge University Press, 1993.

Matthew D Shawkey and Geoffrey E Hill. Carotenoids need structural colours to shine. *Biology letters*, 1(2):121–4, June 2005. ISSN 1744-9561. doi: 10.1098/rsbl.2004.0289.

Jianhong Shen and Yoon-Mo Jung. Weberized Mumford-Shah Model with Bose-Einstein Photon Noise. *Applied Mathematics and Optimization*, 53(3):331–358, March 2006. ISSN 0095-4616. doi: 10.1007/s00245-005-0850-1.

F. Shmid. The Colour Circles by Moses Harris. *The Art Bulletin*, 30(3):227–230, 1948.

Hannah E. Smithson, Greti Dinkova-Bruun, Giles E. M. Gasper, Mike Huxtable, Tom C. B. McLeish, and Cecilia Panti. A three-dimensional color space from the 13th century. *J. Opt. Soc. Am. A*, 29:A346–A352, 2012.

D G Stavenga, R P Smits, and B J Hoenderst. Simple Exponential Functions Describing the Absorbance Bands of Visual Pigment Spectra. *Vision Research*, 33(8):1011 – 1017, 1993.

Doekele G Stavenga, Hein L Leertouwer, N Justin Marshall, and Daniel Osorio. Dramatic colour changes in a bird of paradise caused by uniquely structured breast feather barbules. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715):2098–104, July 2010. ISSN 1471-2954. doi: 10.1098/rspb.2010.2293.

A Stockman and L.T. Sharpe. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, 40(13):1711–1737, 2000.

K Summers and M E Clough. The evolution of coloration and toxicity in the poison frog family (Dendrobatidae). *Proceedings of the National Academy of Sciences of the United States of America*, 98(11):6227–32, May 2001. ISSN 0027-8424. doi: 10.1073/pnas.101134898.

W. A. Sutherland. *Introduction to Metric and Topological Spaces*. 1975.

M Takahashi, H Arita, M Hiraiwahasegawa, and T Hasegawa. Peahens do not prefer peacocks with more elaborate trains. *Animal Behaviour*, 75(4):1209–1219, April 2008. ISSN 00033472. doi: 10.1016/j.anbehav.2007.10.004.

Alfed Tarski and J. H. Woodger. *Logic, Semantics and Metamathematics (J.H.Woodger trans.)*. 1938.

Evan Thompson. *Colour Vision: A Study in Cognative Science and the Philosophy of Perception*. Routledge, 1995.

Flemming Topsœ. Some Inequalities for Information Divergence and Related Measures of Discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

H J Trussell. Applications of Set Theoretic Methods to Color Systems. *Color Research and Application*, 16(1):31–41, 1991.

Constantino Tsallis. Entropic nonextensivity: a possible measure of complexity. *Chaos, Solitons and Fractals*, 13(3):371–391, 2002.

J. A. C. Uy. Modification of the visual background increases the conspicuousness of golden-collared manakin displays. *Behavioral Ecology*, 15(6):1003–1010, June 2004. ISSN 1465-7279. doi: 10.1093/beheco/arh106.

G Von der Emde and B Ronacher. Von der Emde. *J. Comp Physiol A*, 175(6):801–812, 1994.

Hermann von Helmholtz. The Origin and Meaning of Geometrical Axioms. *Mind*, 1(3): 301–321, 1876.

Hermann von Helmholtz. *Handbuch der Physiologishen Optik*. Voss, Hamberg, 2nd edition, 1896.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 1944.

M. Vorobyev, N. Hempel de Ibarra, R. Brandt, and M. Giurfa. Do "White" and "Green" Look the Same to a Bee? *Naturwissenschaften*, 86:592–594, 1999.

Misha Vorobyev. Coloured oil droplets enhance colour discrimination. *Proc. R. Soc. Lond. B*, 270:1255–1261, 2003.

Misha Vorobyev and Daniel Osorio. Receptor noise as a determinant of colour thresholds. *Proc R. Soc. Lond. B*, 265:351 – 358, 1998.

J. J. Vos and P. L. Walraven. An analytical description of the line element in the zone fluctuation model of colour vision. II. The derivation of the line element. *Vision Research*, 12:1345–1365, 1972a.

J. J. Vos and P. L. Walraven. An analytical description of the line element in the zone fluctuation model of colour vision. I. Basic concepts. *Vision Research*, 12:1327–1344, 1972b.

Günter P Wagner and Lee Altenberg. Complex Adaptations and the Evolution of Evolvability. *Evolution*, 50(3):967–976, 1996.

Andreas Weber and Francisco J Varela. Life after Kant : Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, 1:97–125, 2002.

Rüdiger Wehner. 'Matched filters' - neural models of the external world. *J Comp Physiol A*, 161:511–531, 1987.

Rüdiger Wehner. The Hymenopteran Skylight Compas: Matched Filtering and Parallel Coding. *J. exp. Biol*, 146:63–85, 1989.

V. L. Welch and J.-P. Vigneron. Beyond butterfliesthe diversity of biological photonic crystals. *Optical and Quantum Electronics*, 39(4-6):295–303, July 2007. ISSN 0306-8919. doi: 10.1007/s11082-007-9094-4.

Wolfgang Welsh. Animal Aesthetics. *Contemporary Aesthetics*, 2, 2004.

R. Wittkower. Brunelleschi and 'Proportion in Perspective'. *Journal of the Warburg and Courtauld Institutes*, 16(3/4):275–291, 1953.

Günther Wyszecki and W S Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley-Interscience, 2000.

T. Young. The Bakerian Lecture: On the Theory of Light and Colours. *Philosophical Transactions of the Royal Society of London*, 92:12–48, January 1802. ISSN 0261-0523. doi: 10.1098/rstl.1802.0004.

L. A. Zadeh. Fuzzy Algorithms. *Information and Control*, 12:94–102, 1968.

Amotz Zahavi, Avishag Zahavi, Amir Balaban, and Melvin Patrick Ely. *The handicap principle: a missing piece of Darwin's puzzle*. Oxford University Press, 1999.

Jian Zi, Xindi Yu, Yizhou Li, Xinhua Hu, Chun Xu, Xingjun Wang, Xiaohan Liu, and Rongtang Fu. Coloration strategies in peacock feathers. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22):12576–8, October 2003. ISSN 0027-8424. doi: 10.1073/pnas.2133313100.

Appendix A

Notation

A.1 Notation and Conventions

There are a number of specific symbols as well a some conventions for the notation in this document.

A.2 Specific Objects

The following are symbols that have a specific meaning, although they may be overload in some contexts:

Symbol	Meaning
$H(\cdot)$	Entropy
$D^{(KL)}\left(\cdot \,||\, \cdot\right)$	Asymmetric Kullback-Leibler Divergence
$D^{(\alpha)}\left(\cdot \,||\, \cdot\right)$	α-divergence
$D^{(f)}\left(\cdot \,||\, \cdot\right)$	f-divergence
$\mathcal{R}_{\phi}\left(\cdot\right)$	Risk
$\mathcal{R}_{\phi}^{*}$	Optimal Risk
λ	Wavelength
w	Spectral Sensitivity Function
g_{ij}	Indexed Metric Tensor
Γ_{jk}^{i}	Cristoffel Symbol
∂_i (subscript index)	Equivalent to $\frac{\partial}{\partial \xi^i}$
∂	Equivalent to $\frac{\partial}{\partial \xi}$
∂_τ (subscript coordinate)	Equivalent to $\frac{\partial}{\partial \tau}$
$[\cdot]$	Matrix of...
$\mathbb{E}_{\xi}\left[\cdot\right]$	Expectation of ... at point ξ
ℓ	Log probability density

Table A.1: Special Symbols

A.2.1 Common objects

There are a number of conventions that I have tried to stick to throughout, it is not always possible, so the following is in effect a set of 'guidelines' which I have done my best to stick to. I use Latin letters f, g, h for general functions, $c, k, A, B, C \ldots$ are used for constants.

Sets

I write sets in calligraphic script, such as:

$$\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{X} \tag{A.2.1}$$

I will often describe them using set builder notation:

$$\{ \text{ elements} : \text{condition } \} \tag{A.2.2}$$

and $\varnothing$ denotes the empty set.

Probability Theory

I use $\Pr(\cdot)$ to represent probabilities and $p(\cdot)$ or $q(\cdot)$ to represent probability densities. I will use Latin letters for the elements of the support, with the preference:

$$x, y, z, a, b, c \ldots \tag{A.2.3}$$

The corresponding supports are sets with the same letters

$$\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{A}, \mathcal{B}, \mathcal{C} \ldots \tag{A.2.4}$$

I use σ for standard deviations and avoid using μ for the mean.

Measures

I use Σ to denote a σ-algebra, E and element of it (event), and μ and λ as measures.

Coordinates on a Differential Manifold

I use Greek letters for specific coordinates on a differential manifold, with preference left to right:

$$\xi, \rho, \zeta \ldots \tau, \sigma, \mu, \gamma \tag{A.2.5}$$

and the upper case version for the set of all possible coordinates:

$$\Xi, P, Z, \ldots T, \Sigma, M \tag{A.2.6}$$

I use γ for curves (but also for discriminant functions). I use the following letters as indices (in order of preference):

$$i, j, k, l, a, b, c, d, \alpha, \beta \ldots \tag{A.2.7}$$

A.3 Einstein Notation

Einstein notation is used throughout this thesis. Often we wish to sum products of a number of variables, for example, we might wish to calculate an inner product between vectors X and Y, which would be:

$$X_1 Y_1 + X_2 Y_2 + X_3 Y_3 \ldots \tag{A.3.1}$$

or

$$\sum_i X_i Y_i \tag{A.3.2}$$

The power of linear algebra lies in the ability to write this succinctly:

$$XY^T \tag{A.3.3}$$

however, linear algebra becomes problematic. It only works with vectors or matrices: We can have at most two dimensional arrays of objects.

In Einstein's notation there are two types of index: covariant, as in the i in x_i and contravariant, as in the i in x^i. The distinction usually reflects certain geometric properties, but also gives a simple way of writing summations. This notation is centred around inner products, and makes such things very easy to write - when working in the geometric setting for which it is intended, it works very well.

Applying it is pretty simple, when terms with the same index in different locations are multiplied together - assume a sum. Some examples:

$$x^i y^i \rightarrow x^i y^i \tag{A.3.4}$$

$$x^i y_i \rightarrow \sum_i x^i y_i \tag{A.3.5}$$

$$a^i b^j c_{ij} \rightarrow \sum_i \sum_j a^i b^i c_{ij} \tag{A.3.6}$$

$$a^i b^j c^k d_{ijk} \rightarrow \sum_i \sum_j \sum_k a^i b^j c^k d_{ijk} \tag{A.3.7}$$

$$a^i b^j c_k d^k_{ij} \rightarrow \sum_i \sum_j \sum_k a^i b^j c_k d^k_{ij} \tag{A.3.8}$$

$$a^{ij} b_{ij} \rightarrow \sum_i \sum_j a^{ij} b_{ij} \tag{A.3.9}$$

A.3.1 Bracketed Indices

When an index is bracketed, as in $x_{(i)}$ this is to be treated like either upper or lower index in as much as summation should be assumed no matter what its relation is (in terms of upper and lower) to the corresponding indices.

A.3.2 Special Vector Constants

There a number of special objects that will be used throughout, firstly there are two 'constants' $\mathbb{1}$ and $\mathbb{0}$. Which are defined as 1 or 0 respectively, for whatever choice of indexing used. Often it will be used as shorthand for:

$$\mathbb{1}_i x^i = \sum_i x^i \tag{A.3.10}$$

A.3.3 Generalised Delta

I will also use a generalisation of the Kronecker delta function δ. This is defined as:

$$\delta^{b_1 b_2 \ldots ; d_1 d_2 \ldots ; f_1 f_2 \ldots ; \ldots}_{a_1 a_2 \ldots ; c_1 c_2 \ldots ; e_1 e_2 \ldots ; \ldots} = \begin{cases} 1, & \begin{cases} \forall a_i b_j : a_i = b_j \quad \text{and} \\ \forall c_i d_j : c_i = d_j \quad \text{and} \\ \forall e_i f_j : e_i = f_j \quad \text{and} \\ etc \end{cases} \\ 0, & otherwise \end{cases} \tag{A.3.11}$$

The use of the semicolon distinguishes between the logical form[1] $((i = j) \wedge (k = l))$ represented by $ij;kl$ and $(i = j = k = l)$ represented by $ijkl$. In the former it does not necessarily hold that, for example $j = k$.

For this δ, the following two identities hold. Firstly we can concatenate indices (separating by ";") when two δs are multiplied

$$\left(\delta^{b_1 b_2 \ldots ; d_1 d_2 \ldots ; \ldots}_{a_1 a_2 \ldots ; c_1 c_2 \ldots ; \ldots}\right) \left(\delta^{f_1 f_2 \ldots ; h_1 h_2 \ldots ; \ldots}_{e_1 e_2 \ldots ; g_1 g_2 \ldots ; \ldots}\right) = \delta^{b_1 b_2 \ldots ; d_1 d_2 \ldots ; f_1 f_2 \ldots ; h_1 h_2 \ldots ; \ldots}_{a_1 a_2 \ldots ; c_1 c_2 \ldots ; e_1 e_2 \ldots ; g_1 g_2 \ldots ; \ldots} \tag{A.3.12}$$

and, we can remove a ";" by contracting across it, here the contracting index is x.

$$\delta^{b_1 b_2 \ldots ;\, \mathbf{x} d_1 d_2 \ldots ; \ldots}_{\mathbf{x} a_1 a_2 \ldots ;\, c_1 c_2 \ldots ; \ldots} = \delta^{b_1 b_2 \ldots d_1 d_2 \ldots ; \ldots}_{a_1 a_2 \ldots c_1 c_2 \ldots ; \ldots} \tag{A.3.13}$$

An example application of these identities:

$$\delta^{al}_{ij} \delta^{b}_{kl} = \delta^{a\mathbf{l};\, b}_{ij;\, k\mathbf{l}} \tag{A.3.14}$$

$$= \delta^{ab}_{ijk} \tag{A.3.15}$$

A common use of this delta function is in situations like:

$$\delta_{abc} x^a y^b z^c = \sum_i x^i y^i z^i \tag{A.3.16}$$

[1] $\wedge$ represents logical conjunction - "AND"

Appendix B

Binary Choice and Linear Discriminants

We have a score for behaviours (ω) and (α): $s(\alpha, \omega)$. We wish to show that maximising the expectation (over a Borel set) of this score given two probability distributions ($\Pr(x,\ \alpha) = \Pr(\alpha)\Pr(x \mid \alpha),\ \alpha = \alpha_1, \alpha_2$) gives a choice function where:

$$\omega = \begin{cases} \omega_1, & \Pr(x,\ \alpha_1) > k \Pr(x,\ \alpha_2) \\ \omega_2, & \textit{otherwise} \end{cases} \tag{B.0.1}$$

where

$$k = \frac{s(2,2) - s(2,1)}{s(1,1) - s(1,2)} \tag{B.0.2}$$

The total score for a choice (determined by a proposition ϕ) can be written as

$$\begin{aligned} S(\phi) &= \int \bar{s}_\phi(x) \mathrm{d}x & \text{(B.0.3)} \\ &= s(1,1) \int_\Phi \Pr(x,\ \alpha_1) \mathrm{d}x \\ &\quad + s(1,2) \int_\Psi \Pr(x,\ \alpha_1) \mathrm{d}x \\ &\quad + s(2,1) \int_\Phi \Pr(x,\ \alpha_2) \mathrm{d}x \\ &\quad + s(2,1) \int_\Psi \Pr(x,\ \alpha_2) \mathrm{d}x & \text{(B.0.4)} \end{aligned}$$

where $\phi(x) = \neg\psi(x)$ are propositions about x, which go on to define the disjoint sets

$$\Phi = \{x\ :\ \phi(x)\} \ \text{ and } \ \Psi = \{x\ :\ \psi(x)\} \tag{B.0.5}$$

A corollary of which is that $x \in \Phi \implies \omega = \omega_1$ and $x \in \Psi \implies \omega = \omega_2$. We can maximise S in a point-wise manner, by considering the two possibilities for a particular x:

$$\begin{aligned}
\phi(x) = \texttt{true} \quad &: \quad \bar{s}_\phi(x) = s(1,1)\Pr(x,\ \alpha_1) + s(2,1)\Pr(x,\ \alpha_2) \\
\phi(x) = \texttt{false} \quad &: \quad \bar{s}_\phi(x) = s(1,2)\Pr(x,\ \alpha_1) + s(2,2)\Pr(x,\ \alpha_2)
\end{aligned}$$

So, one just chooses the bigger one if we want to maximise S or the smaller if one wants to minimise. For maximisation:

$$\begin{aligned}
\phi(x) \quad &= \quad [\![s(1,1)\Pr(x,\ \alpha_1) + s(2,1)\Pr(x,\ \alpha_2) \\
&\qquad > s(1,2)\Pr(x,\ \alpha_1) + s(2,2)\Pr(x,\ \alpha_2)]\!] \qquad \text{(B.0.6)} \\
&= \quad [\![(s(1,1) - s(1,2))\Pr(x,\ \alpha_1) \\
&\qquad > (s(2,2) - s(2,1))\Pr(x,\ \alpha_2)]\!] \qquad \text{(B.0.7)}
\end{aligned}$$

and assuming $s(1,1) > s(1,2)$:

$$\phi(x) = [\![\Pr(x,\ \alpha_1) > \frac{s(2,2) - s(2,1)}{s(1,1) - s(1,2)} \Pr(x,\ \alpha_2)]\!] \qquad \text{(B.0.8)}$$

as required.

Appendix C

Divergences

This appendix contains the supplementary material associated with calculations involving divergences.

C.1 Expansion of f-divergences

The following is a straight forwards Taylor expansion of f-divergences. Due to it's length and triviality it has not been reported in the literature. Performing the expansion and checking the results for errors does however require a significant amount of work - even with infernal newfangled contraptions capable of advanced symbolic manipulation. For this reason it is included here. The f-divergence:

$$D^{(f)}\left(\xi \mid\mid \rho\right) = \int_{\chi} p_{\xi} f\left(\frac{p_{\rho}}{p_{\xi}}\right) \mathrm{d}\mu(x) \tag{C.1.1}$$

will be expanded as (see 4.5.1):

$$\begin{aligned} D^{(h)}\left(\xi \mid\mid \xi + \Delta\xi\right) \;=\; & D^{(h)}\left(\xi \mid\mid \rho\right)\Big|_{\rho=\xi} + \\ & \frac{\partial}{\partial\rho^{i}} D^{(h)}\left(\xi \mid\mid \rho\right)\Big|_{\rho=\xi} \Delta\xi^{i} + \\ & \frac{1}{2}\frac{\partial}{\partial\rho^{i}}\frac{\partial}{\partial\rho^{j}} D^{(h)}\left(\xi \mid\mid \rho\right)\Big|_{\rho=\xi} \Delta\xi^{i}\Delta\xi^{j} + \\ & \frac{1}{6}\frac{\partial}{\partial\rho^{i}}\frac{\partial}{\partial\rho^{j}}\frac{\partial}{\partial\rho^{k}} D^{(h)}\left(\xi \mid\mid \rho\right)\Big|_{\rho=\xi} \Delta\xi^{i}\Delta\xi^{j}\Delta\xi^{k} + \\ & o(\Delta\xi^{4}) \end{aligned} \tag{C.1.2}$$

assuming continuity (to third order) of $f(u)$ at $u = 1$. Here there is no other assumption about the nature of f. The following uses non-normalised probabilities, with

$$\int_{\chi} p_{\xi} \mathrm{d}\mu(x) = \tau \tag{C.1.3}$$

which allows for a more general result. We also need to recognise that:

$$\begin{aligned}\frac{\partial}{\partial\rho^i}f\left(\frac{p_\rho}{p_\xi}\right) &= \frac{\partial}{\partial\rho^i}f(u)\\ &= \frac{\partial u}{\partial\rho^i}f'(u)\\ &= \left(\frac{1}{p_\xi}\right)\left(\frac{\partial}{\partial\rho^i}p_\rho\right)f'(u) \qquad \text{(C.1.4)}\end{aligned}$$

which is used throughout. As usual, I will use the notation of $\ell_\xi = \log p_\xi$ and $\partial_i = \frac{\partial}{\partial\xi^i}$.

From here I will take each term of the expansion shown in C.1.2 in order.

Zeroth order in $\Delta\xi$

$$\begin{aligned}D^{(f)}(\,\xi\,||\,\rho\,)\Big|_{\rho=\xi} &= \int_\chi p_\xi f\left(\frac{p_\xi}{p_\xi}\right)\mathrm{d}\mu(x) \qquad \text{(C.1.5)}\\ &= f(1)\int_\chi p_\xi \mathrm{d}\mu(x) \qquad \text{(C.1.6)}\\ &= \tau f(1) \qquad \text{(C.1.7)}\end{aligned}$$

First order in $\Delta\xi$

First the differential, beginning with the product rule and noting $\frac{\partial}{\partial\rho^i}p_\xi = 0$:

$$\begin{aligned}\frac{\partial}{\partial\rho^i}D^{(f)}(\,\xi\,||\,\rho\,) &= \int_\chi\left(\frac{\partial}{\partial\rho^i}p_\xi\right)f\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) +\\ &\quad \int_\chi p_\xi\frac{\partial}{\partial\rho^i}f\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) \qquad \text{(C.1.8)}\\ &= \int_\chi\left(\frac{\partial}{\partial\rho^i}p_\rho\right)f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) \qquad \text{(C.1.9)}\end{aligned}$$

evaluating at $\rho = \xi$:

$$\begin{aligned}\frac{\partial}{\partial\rho^i}D^{(f)}(\,\xi\,||\,\rho\,)\Big|_{\rho=\xi} &= \int_\chi\left(\frac{\partial}{\partial\xi^i}p_\xi\right)f'(1)\,\mathrm{d}\mu(x) \qquad \text{(C.1.10)}\\ &= f'(1)\int_\chi\frac{\partial}{\partial\xi^i}p_\xi\mathrm{d}\mu(x) \qquad \text{(C.1.11)}\\ &= f'(1)\partial_i\tau \qquad \text{(C.1.12)}\end{aligned}$$

Second order in $\Delta\xi$

Beginning with equation C.1.9:

$$\begin{aligned}\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}D^{(f)}(\,\xi\,||\,\rho\,) &= \frac{\partial}{\partial\rho^j}\int_\chi\left(\frac{\partial}{\partial\rho^i}p_\rho\right)f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)\\ &= \int_\chi\left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}p_\rho\right)f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) +\\ &\quad \int_\chi\left(\frac{\partial}{\partial\rho^i}p_\rho\right)\frac{\partial}{\partial\rho^j}f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)\end{aligned} \qquad \text{(C.1.13)}$$

$$
\begin{aligned}
&= \int_\chi \left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}p_\rho\right) f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) + \\
&\quad \int_\chi \frac{1}{p_\xi}\left(\frac{\partial}{\partial\rho^i}p_\rho\right)\left(\frac{\partial}{\partial\rho^j}p_\rho\right) f''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)
\end{aligned}
\tag{C.1.14}
$$

evaluating at $\rho = \xi$:

$$
\begin{aligned}
\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}D^{(f)}\left(\,\xi\,||\,\rho\,\right)\Big|_{\rho=\xi} &= f'(1)\int_\chi \frac{\partial}{\partial\xi^i}\frac{\partial}{\partial\xi^j}p_\xi \mathrm{d}\mu(x) \\
&+ f''(1)\int_\chi \left(\frac{1}{p_\xi}\frac{\partial}{\partial\xi^i}p_\xi\right)\left(\frac{1}{p_\xi}\frac{\partial}{\partial\xi^j}p_\xi\right)p_\xi \mathrm{d}\mu(x) \\
&= f'(1)\partial_i\partial_j\tau + f''(1)\int_\chi \left(\frac{\partial}{\partial\xi^i}\ell_\xi\right)\left(\frac{\partial}{\partial\xi^j}\ell_\xi\right)p_\xi \mathrm{d}\mu(x) \\
&= f'(1)\partial_i\partial_j\tau + f''(1)\tau g_{ij}
\end{aligned}
\tag{C.1.15}
$$

where g_{ij} is the Fisher metric (we must remember that expectations are taken over normalised distributions):

$$
\begin{aligned}
g_{ij} &= \mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\right]\mathrm{d}\mu(x) \\
&= \int_\chi \left(\frac{\partial}{\partial\xi^i}\ell_\xi\right)\left(\frac{\partial}{\partial\xi^j}\ell_\xi\right)\frac{p_\xi}{\int_\chi p_\xi \mathrm{d}\mu(x)}\mathrm{d}\mu(x) \\
&= \frac{1}{\tau}\int_\chi \left(\frac{\partial}{\partial\xi^i}\ell_\xi\right)\left(\frac{\partial}{\partial\xi^j}\ell_\xi\right)p_\xi \mathrm{d}\mu(x)
\end{aligned}
$$

Third order in $\Delta\xi$

Beginning with equation C.1.14:

$$
\begin{aligned}
&\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}\frac{\partial}{\partial\rho^k}D^{(f)}\left(\,\xi\,||\,\rho\,\right) = \\
&\quad \frac{\partial}{\partial\rho^k}\int_\chi \left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}p_\rho\right) f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) \\
&+ \quad \frac{\partial}{\partial\rho^k}\int_\chi \frac{1}{p_\xi}\left(\frac{\partial}{\partial\rho^i}p_\rho\right)\left(\frac{\partial}{\partial\rho^j}p_\rho\right) f''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)
\end{aligned}
\tag{C.1.16}
$$

Taking each part of the sum separately gives for the left hand summand:

$$
\begin{aligned}
&\frac{\partial}{\partial\rho^k}\int_\chi \left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}p_\rho\right) f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) = \\
&\quad \int_\chi \left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}\frac{\partial}{\partial\rho^k}p_\rho\right) f'\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) \\
&+ \quad \int_\chi \frac{1}{p_\xi}\left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^j}p_\rho\right)\left(\frac{\partial}{\partial\rho^k}p_\rho\right) f''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)
\end{aligned}
\tag{C.1.17}
$$

evaluating at $\rho = \xi$:

$$LH = f'(1)\partial_i\partial_j\partial_k\tau + f''(1)\int_\chi \left(\frac{\partial}{\partial\xi^i}\frac{\partial}{\partial\xi^j}p_\xi\right)\left(\frac{\partial}{\partial\xi^k}\ell_\xi\right)\mathrm{d}\mu(x) \tag{C.1.18}$$

Now, it is possible to write:

$$\int_\chi \left(\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}p_\xi\right)\left(\frac{\partial}{\partial\xi^c}\ell_\xi\right)\mathrm{d}\mu(x)$$

as

$$\int_\chi p_\xi \left(\frac{1}{p_\xi}\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}p_\xi\right)\left(\frac{\partial}{\partial\xi^c}\ell_\xi\right)\mathrm{d}\mu(x)$$

and also, we can see that:

$$\begin{aligned}
\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}\ell_\xi &= \frac{\partial}{\partial\xi^a}\left(\frac{1}{p_\xi}\frac{\partial}{\partial\xi^b}p_\xi\right) \\
&= \frac{-1}{p_\xi^2}\frac{\partial}{\partial\xi^a}p_\xi\frac{\partial}{\partial\xi^b}p_\xi + \frac{1}{p_\xi}\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}p_\xi \\
&= -\frac{\partial}{\partial\xi^a}\ell_\xi\frac{\partial}{\partial\xi^b}\ell_\xi + \frac{1}{p_\xi}\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}p_\xi
\end{aligned} \tag{C.1.19}$$

so

$$\frac{1}{p_\xi}\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}p_\xi = \frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}\ell_\xi + \frac{\partial}{\partial\xi^a}\ell_\xi\frac{\partial}{\partial\xi^b}\ell_\xi \tag{C.1.20}$$

and

$$\int_\chi p_\xi \left(\frac{1}{p_\xi}\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}p_\xi\right)\left(\frac{\partial}{\partial\xi^c}\ell_\xi\right)\mathrm{d}\mu(x)$$

becomes

$$\begin{aligned}
&\int_\chi p_\xi\frac{\partial}{\partial\xi^a}\frac{\partial}{\partial\xi^b}\ell_\xi\frac{\partial}{\partial\xi^c}\ell_\xi\mathrm{d}\mu(x) \\
+ &\int_\chi p_\xi\frac{\partial}{\partial\xi^a}\ell_\xi\frac{\partial}{\partial\xi^b}\ell_\xi\frac{\partial}{\partial\xi^c}\ell_\xi\mathrm{d}\mu(x)
\end{aligned} \tag{C.1.21}$$

which can be written as the sum of expectations:

$$\tau\mathbb{E}_\xi\left[\partial_a\partial_b\ell_\xi\partial_c\ell_\xi\right] + \tau\mathbb{E}_\xi\left[\partial_a\ell_\xi\partial_b\ell_\xi\partial_c\ell_\xi\right] \tag{C.1.22}$$

This means equation C.1.18 can be written:

$$\begin{aligned}
LH &= f'(1)\partial_i\partial_j\partial_k\tau + \\
&\quad \tau f''(1)\mathbb{E}_\xi\left[\partial_i\partial_j\ell_\xi\partial_k\ell_\xi\right] + \\
&\quad \tau f''(1)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]
\end{aligned} \tag{C.1.23}$$

Now, Taking the right hand part of the sum in C.1.16:

$$\frac{\partial}{\partial\rho^k}\int_\chi \frac{1}{p_\xi}\left(\frac{\partial}{\partial\rho^i}p_\rho\right)\left(\frac{\partial}{\partial\rho^j}p_\rho\right)f''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x) =$$

$$\begin{aligned}
&\int_\chi \frac{1}{p_\xi}\left(\frac{\partial}{\partial\rho^i}\frac{\partial}{\partial\rho^k}p_\rho\right)\left(\frac{\partial}{\partial\rho^j}p_\rho\right)f''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)\\
+\ &\int_\chi \frac{1}{p_\xi}\left(\frac{\partial}{\partial\rho^i}p_\rho\right)\left(\frac{\partial}{\partial\rho^j}\frac{\partial}{\partial\rho^k}p_\rho\right)f''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)\\
+\ &\int_\chi \frac{1}{p_\xi^2}\left(\frac{\partial}{\partial\rho^i}p_\rho\right)\left(\frac{\partial}{\partial\rho^j}p_\rho\right)\left(\frac{\partial}{\partial\rho^k}p_\rho\right)f'''\left(\frac{p_\rho}{p_\xi}\right)\mathrm{d}\mu(x)
\end{aligned} \tag{C.1.24}$$

evaluating at $\rho=\xi$:

$$\begin{aligned}
RH \ =\ & f''(1)\int_\chi\left(\frac{\partial}{\partial\xi^j}\frac{\partial}{\partial\xi^k}p_\xi\right)\left(\frac{\partial}{\partial\xi^i}\ell_\xi\right)\mathrm{d}\mu(x)+ && \text{(C.1.25)}\\
& f''(1)\int_\chi\left(\frac{\partial}{\partial\xi^i}\frac{\partial}{\partial\xi^k}p_\xi\right)\left(\frac{\partial}{\partial\xi^j}\ell_\xi\right)\mathrm{d}\mu(x)+ && \text{(C.1.26)}\\
& \tau f'''(1)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right] && \text{(C.1.27)}
\end{aligned}$$

which can also be completely written in terms of expectations:

$$\begin{aligned}
RH \ =\ & \tau f''(1)\mathbb{E}_\xi\left[\partial_j\partial_k\ell_\xi\partial_i\ell_\xi\right]+\\
& \tau f''(1)\mathbb{E}_\xi\left[\partial_i\partial_k\ell_\xi\partial_j\ell_\xi\right]+\\
& 2\tau f''(1)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]+\\
& \tau f'''(1)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]
\end{aligned} \tag{C.1.28}$$

adding the left and right sides gives:

$$\begin{aligned}
LH+RH \ =\ & f'(1)\partial_i\partial_j\partial_k\tau+\\
& \tau f''(1)\mathbb{E}_\xi\left[\partial_i\partial_j\ell_\xi\partial_k\ell_\xi\right]+\\
& \tau f''(1)\mathbb{E}_\xi\left[\partial_j\partial_k\ell_\xi\partial_i\ell_\xi\right]+\\
& \tau f''(1)\mathbb{E}_\xi\left[\partial_i\partial_k\ell_\xi\partial_j\ell_\xi\right]+\\
& 3\tau f''(1)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]+\\
& \tau f'''(1)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]
\end{aligned} \tag{C.1.29}$$

Next, we note that:

$$\begin{aligned}
\partial_k g_{ij} \ &=\ \partial_k\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\right]\\
&=\ \mathbb{E}_\xi\left[\partial_i\partial_k\ell_\xi\partial_j\ell_\xi\right]+\\
&=\ \mathbb{E}_\xi\left[\partial_j\partial_k\ell_\xi\partial_i\ell_\xi\right]+\\
&=\ \mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]
\end{aligned} \tag{C.1.30}$$

so

$$\begin{aligned} LH + RH &= f'(1)\partial_i\partial_j\partial_k\tau + \tau f''(1)\partial_k g_{ij} + \\ &\quad \tau f''(1)\mathbb{E}_\xi\left[\partial_i\partial_j\ell_\xi\partial_k\ell_\xi\right] + \\ &\quad \tau(2f''(1) + f'''(1))\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right] \\ &= f'(1)\partial_i\partial_j\partial_k\tau + \tau f''(1)\partial_k g_{ij} + \\ &\quad \tau f''(1)\Bigg(\mathbb{E}_\xi\left[\partial_i\partial_j\ell_\xi\partial_k\ell_\xi\right] + \\ &\quad \left(2 + \frac{f'''(1)}{f''(1)}\right)\mathbb{E}_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right]\Bigg) \end{aligned}$$

$$\begin{aligned} &= f'(1)\partial_i\partial_j\partial_k\tau + \tau f''(1)\partial_k g_{ij} + \\ &\quad \tau f''(1)\mathbb{E}_\xi\left[\left(\partial_i\partial_j\ell_\xi + \left(2 + \frac{f'''(1)}{f''(1)}\right)\partial_i\ell_\xi\partial_j\ell_\xi\right)(\partial_k\ell_\xi)\right] \end{aligned} \tag{C.1.31}$$

remembering that the definition of the Cristoffel symbols of the α-connection are given by:

$$\Gamma^{(\alpha)}_{ij,k} = \mathbb{E}_\xi\left[\left(\partial_i\partial_j\ell_\xi + \frac{1-\alpha}{2}\partial_i\ell_\xi\partial_j\ell_\xi\right)(\partial_k\ell_\xi)\right] \tag{C.1.32}$$

we see that if we let $-\alpha = 3 + 2f'''(1)/f''(1)$ then:

$$LH + RH = f'(1)\partial_i\partial_j\partial_k\tau + \tau f''(1)\partial_k g_{ij} + \tau f''(1)\Gamma^{-(3+2f'''(1)/f''(1))}_{ij,k} \tag{C.1.33}$$

Final Expansion

$$\begin{aligned} D^{(f)}\left(\,\xi \,||\, \xi + \Delta\xi\,\right) &= \tau f(1) + \\ &\quad f'(1)\partial_i\tau\Delta\xi^i + \\ &\quad \tfrac{1}{2}f'(1)\partial_i\partial_j\tau\Delta\xi^i\Delta\xi^j + \\ &\quad \tfrac{1}{2}\tau f''(1)g_{ij}\Delta\xi^i\Delta\xi^j + \\ &\quad \tfrac{1}{6}f'(1)\partial_i\partial_j\partial_k\tau\Delta\xi^i\Delta\xi^j\Delta\xi^k + \\ &\quad \tfrac{1}{6}\tau f''(1)\partial_k g_{ij}\Delta\xi^i\Delta\xi^j\Delta\xi^k + \\ &\quad \tfrac{1}{6}\tau f''(1)\Gamma^{-(3+2f'''(1)/f''(1))}_{ij,k}\Delta\xi^i\Delta\xi^j\Delta\xi^k + \\ &\quad o(\Delta\xi^4) \end{aligned} \tag{C.1.34}$$

This can be cleaned up by subtracting $f'(1)\tau(\xi + \Delta\xi)$ from the expanded function:

$$D^{(f)}\left(\,\xi \,||\, \xi + \Delta\xi\,\right) - f'(1)\tau(\xi + \Delta\xi) =$$

$$
\begin{aligned}
&\tau f(1) + \\
&\tfrac{1}{2}\tau f''(1) g_{ij} \Delta\xi^i \Delta\xi^j + \\
&\tfrac{1}{6}\tau f''(1) \partial_k g_{ij} \Delta\xi^i \Delta\xi^j \Delta\xi^k + \\
&\tfrac{1}{6}\tau f''(1) \Gamma^{-(3+2f'''(1)/f''(1))}_{ij,k} \Delta\xi^i \Delta\xi^j \Delta\xi^k + \\
&o(\Delta\xi^4)
\end{aligned}
\tag{C.1.35}
$$

h-divergences

For an h-divergence we have $f(1) = f'(1) = 0$ so:

$$
\begin{aligned}
D^{(h)}\left(\,\xi \,||\, \xi + \Delta\xi\,\right) \;=\; & \tfrac{1}{2}\tau f''(1) g_{ij} \Delta\xi^i \Delta\xi^j + \\
& \tfrac{1}{6}\tau f''(1) \left(\partial_k g_{ij} + \Gamma^{-(3+2f'''(1)/f''(1))}_{ij,k}\right) \Delta\xi^i \Delta\xi^j \Delta\xi^k + \\
& o(\Delta\xi^4)
\end{aligned}
\tag{C.1.36}
$$

C.2 Preservation of Symmetry under the Geometrising Transform

This appendix demonstrates that the geometrising transform

$$
f(u) \to f(u) + (1-u) f'(1) - f(1) \tag{C.2.1}
$$

preserves the relationship (symmetry)

$$
D^{(f)}\left(\,\xi \,||\, \rho\,\right) = D^{(f)}\left(\,\rho \,||\, \xi\,\right) \tag{C.2.2}
$$

in the general case of different prior probabilities.

Let us consider two possibilities ξ and ρ with prior probabilities $\pi^\xi = \Pr(\xi)$ and $\pi^\rho = \Pr(\rho)$ and distributions conditional on them, $\eta^\xi = \Pr(x \,|\, \xi)$ and $\eta^\rho = \Pr(x \,|\, \rho)$. The f-divergences corresponding to the difference between the two outcomes - prior probabilities considered - is given by:

$$
D^{(f)}\left(\,\xi \,||\, \rho\,\right) = \int_{\mathcal{X}} \pi^\rho \eta^\rho f\left(\frac{\pi^\xi \eta^\xi}{\pi^\rho \eta^\rho}\right) \mathrm{d}\mu(x) \tag{C.2.3}
$$

Letting h be the transformed function $h(u) = f(u) + (1-u)f'(1) - f(1)$ then:

$$\begin{aligned} & D^{(h)}\left(\xi \mid\mid \rho\right) \\ = & \int_{\mathcal{X}} \pi^{\rho}\eta^{\rho}\left(f\left(\frac{\pi^{\xi}\eta^{\xi}}{\pi^{\rho}\eta^{\rho}}\right) + \left(1 - \frac{\pi^{\xi}\eta^{\xi}}{\pi^{\rho}\eta^{\rho}}\right)f'(1) - f(1)\right)\mathrm{d}\mu(x) \\ = & \int_{\mathcal{X}} \pi^{\rho}\eta^{\rho} f\left(\frac{\pi^{\xi}\eta^{\xi}}{\pi^{\rho}\eta^{\rho}}\right) + (\pi^{\rho}\eta^{\rho} - \pi^{\xi}\eta^{\xi})f'(1) - \pi^{\rho}\eta^{\rho}f(1)\ \ \mathrm{d}\mu(x) \\ = & D^{(f)}\left(\xi \mid\mid \rho\right) + f'(1)\pi^{\xi}\int_{\mathcal{X}} \eta^{\xi}\mathrm{d}\mu(x) \\ & -f'(1)\pi^{\xi}\int_{\mathcal{X}} \eta^{\xi}\mathrm{d}\mu(x) + f(1)\pi^{\rho}\int_{\mathcal{X}} \eta^{\rho}\mathrm{d}\mu(x) \end{aligned}$$

noting that $\int_{\mathcal{X}} \eta^{\xi}\mathrm{d}\mu(x) = \int_{\mathcal{X}} \eta^{\rho}\mathrm{d}\mu(x) = 1$ we have:

$$D^{(h)}\left(\xi \mid\mid \rho\right) = D^{(f)}\left(\xi \mid\mid \rho\right) + (\pi^{\rho} - \pi^{\xi})f'(1) - \pi^{\rho}f(1) \tag{C.2.4}$$

Imposing the symmetry constraint in equation C.2.2 onto this transformed divergence gives:

$$\begin{aligned} 0 &= D^{(h)}\left(\xi \mid\mid \rho\right) - D^{(h)}\left(\xi \mid\mid \rho\right) \\ &= D^{(f)}\left(\xi \mid\mid \rho\right) - D^{(f)}\left(\xi \mid\mid \rho\right) + \left(\pi^{\xi} - \pi^{\rho}\right)\left(2f'(1) - f(1)\right) \end{aligned} \tag{C.2.5}$$

Now, since we know that the relationship holds for $D^{(f)}\left(\xi \mid\mid \rho\right)$ then we have the requirement that:

$$\left(\pi^{\xi} - \pi^{\rho}\right)\left(2f'(1) - f(1)\right) = 0 \tag{C.2.6}$$

we also know that this symmetry implies that the following relationship holds for f:

$$f(u) = uf\left(1/u\right) \tag{C.2.7}$$

Differentiating this:

$$f'(u) = f(1/u) - \frac{f'(1/u)}{u} \tag{C.2.8}$$

and evaluating at $u = 1$:

$$2f'(1) - f(1) = 0 \tag{C.2.9}$$

leads to a condition where by necessity equation C.2.6 holds, therefor the geometrising transform preserves this symmetry.

Appendix D

Specific Values of the Fisher Metric

The following is the derivation of the Fisher metric for some of the distributions used in this thesis. I will show them for the one dimensional case, as generalisation to the multidimensional case is trivial.

D.1 Poisson Distribution

If

$$X \sim \textsc{Poisson}\,(\lambda) \tag{D.1.1}$$

then

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{D.1.2}$$

The log probability is:

$$\ell(x;\lambda) = -\lambda + x\log\lambda - \log x! \tag{D.1.3}$$

and

$$\frac{\partial}{\partial\lambda}\ell(x;\lambda) = \frac{x}{\lambda} - 1 \tag{D.1.4}$$

$$\left(\frac{\partial}{\partial\lambda}\ell(x;\lambda)\right)^2 = \left(\frac{x}{\lambda} - 1\right)^2 \tag{D.1.5}$$

So the Fisher information g is given by:

$$g = \sum_{x=0}^{\infty}\left(\frac{x}{\lambda} - 1\right)^2 \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{\lambda} \tag{D.1.6}$$

D.2 Gamma Distribution

Here I only consider the case of a fixed population size, where k is a constant.

If

$$X \sim \textsc{Gamma}\,(k, \theta) \tag{D.2.1}$$

then

$$\Pr(X = x) = \gamma(x, k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k} \tag{D.2.2}$$

where $\Gamma : \mathbb{C} \to \mathbb{C}$ is the gamma function:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t \tag{D.2.3}$$

for $n \in \mathbb{N}$ we have the relation:

$$\Gamma(n) = (n-1)! \tag{D.2.4}$$

and similarly:

$$\Gamma(x+1) = x\Gamma(x) \tag{D.2.5}$$

which holds for all $x \in \mathbb{R}_+$. Now, we have a statistical manifold with log probability given by:

$$\ell(x; k, \theta) = (k-1)\log x - x/\theta - \log\Gamma(k) - k\log\theta \tag{D.2.6}$$

so

$$\frac{\partial}{\partial\theta}\ell(x; k, \theta) = \frac{x}{\theta^2} - \frac{k}{\theta} \tag{D.2.7}$$

$$\left(\frac{\partial}{\partial\theta}\ell(x; k, \theta)\right)^2 = \frac{1}{\theta^2}\left(\frac{x}{\theta} - k\right)^2 \tag{D.2.8}$$

and the Fisher information g is:

$$\begin{aligned}
g &= \int_{-\infty}^{\infty} \frac{1}{\theta^2}\left(\frac{x}{\theta} - k\right)^2 \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}\mathrm{d}x && \text{(D.2.9)}\\
&= \frac{1}{\theta^2}\int_{-\infty}^{\infty} \left(\frac{x}{\theta} - k\right)^2 \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}\mathrm{d}x && \text{(D.2.10)}\\
&= \frac{1}{\theta^2}\int_{-\infty}^{\infty} \left(\frac{x^2}{\theta^2} - 2\frac{kx}{\theta} + k^2\right) \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}\mathrm{d}x && \text{(D.2.11)}\\
&= \frac{1}{\theta^2}(A - 2B + C) && \text{(D.2.12)}
\end{aligned}$$

where, as $\int_{-\infty}^{\infty} \gamma(x,k,\theta)\mathrm{d}x = 1$:

$$\begin{aligned} A &= \int_{-\infty}^{\infty} \frac{x^2}{\theta^2} \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k} && \text{(D.2.13)} \\ &= \int_{-\infty}^{\infty} k(k+1)\gamma(x,k,\theta)\mathrm{d}x = k(k+1) && \text{(D.2.14)} \\ B &= \int_{-\infty}^{\infty} \frac{kx}{\theta} \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k} = \int_{-\infty}^{\infty} k^2\gamma(x,k,\theta)\mathrm{d}x = k^2 && \text{(D.2.15)} \\ C &= \int_{-\infty}^{\infty} k^2 \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k} = \int_{-\infty}^{\infty} k^2\gamma(x,k,\theta)\mathrm{d}x = k^2 && \text{(D.2.16)} \end{aligned}$$

so

$$A - 2B + C = k(k+1) - 2k^2 + k^2 = k \qquad \text{(D.2.17)}$$

thus:

$$g = \frac{k}{\theta^2} \qquad \text{(D.2.18)}$$

D.3 Binomial Distribution

Here I only consider the case of a fixed population size, where n is a constant.

If

$$X \sim \textsc{Binomial}\,(n,p) \qquad \text{(D.3.1)}$$

then

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{(D.3.2)}$$

so the log probability is:

$$\ell(x;n,p) = \log \binom{n}{x} + x\log p - (n-x)\log(1-p) \qquad \text{(D.3.3)}$$

and

$$\frac{\partial}{\partial p}\ell(x;n,p) = \frac{x - np}{p(1-p)} \qquad \text{(D.3.4)}$$

so that the Fisher information is given by the sum:

$$
\begin{aligned}
g &= \sum_{x=0}^{n} \binom{n}{x} \frac{(x-np)^2}{p^2(1-p)^2} p^x (1-p)^{n-x} && \text{(D.3.5)} \\
&= \sum_{x=0}^{n} \binom{n}{x} (x-np)^2 p^{x-2} (1-p)^{n-x-2} && \text{(D.3.6)} \\
&= \sum_{x=0}^{n} \binom{n}{x} x^2 p^{x-2} (1-p)^{n-x-2} && \text{(D.3.7)} \\
&\quad - \sum_{x=0}^{n} \binom{n}{x} 2npp^{x-2} (1-p)^{n-x-2} && \text{(D.3.8)} \\
&\quad + \sum_{x=0}^{n} \binom{n}{x} n^2 p^2 p^{x-2} (1-p)^{n-x-2} && \text{(D.3.9)}
\end{aligned}
$$

which can be written as expectations of powers of x, then in terms of the mean and variance:

$$
\begin{aligned}
g &= \frac{1}{p^2(1-p)^2} \left(\mathbb{E}_p \left[x^2 \right] - 2np \, \mathbb{E}_p \left[x \right] + n^2 p^2 \right) \right) && \text{(D.3.10)} \\
&= \frac{1}{p^2(1-p)^2} \left(\sigma^2 + (\mu - np)^2 \right) && \text{(D.3.11)}
\end{aligned}
$$

and it is known that the mean of a binomial distribution is np and the variance is $np(1-p)$ so we have:

$$\mu - np = 0 \tag{D.3.12}$$

so

$$g = \frac{\sigma^2}{p^2(1-p)^2} \tag{D.3.13}$$

and finally

$$g = \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)} \tag{D.3.14}$$

Appendix E

Sexual Selection Simulation

E.1 Model

The model of signal production though sexual selection consists of two collections of species, the 'signallers' and the 'drifters', labelled $\mathcal{S}$ and $\mathcal{D}$ respectively. They contain n_S and n_D species so that:

$$\mathcal{S} = \{s_1, s_2 \dots s_{n_S}\}, \qquad \mathcal{D} = \{d_1, d_2 \dots d_{n_D}\} \tag{E.1.1}$$

Each species s_i or d_i contains n_O organisms which have variables corresponding to pigments that define their appearance A and their visual systems V and the colour of their preferred mate T. In addition there is an integer identifier for the organism I, and its species S. The members of each species are each then defined by a tuple (A, V, T, I, S). The appearance, A, is a vector with n_p elements (where n_p is the number of samples that the spectrum is recorded at) and the visual pigment genes V is a n_e vector which defines an n_e-by-n_p matrix M (where n_e is the number of visual pigment types), so that the colour of an organism a as seen by an organism b ($c_b(a)$) is given by the usual inner product:

$$c_b(a) = M(V_b) \cdot A_a \tag{E.1.2}$$

where subscripts denote the organism to which the variables are associated. The function M is given by

$$M_{ij}(V) = \frac{1 + \sin(\pi(\frac{1}{2}i + 2V_j))}{2 + \frac{1}{2}\sum_i \sin(\pi(\frac{1}{2}i + 2V_j))} \tag{E.1.3}$$

Though it is rather arbitrary, it is chosen to have a single peak and symmetry, as well as being normalised (the normalising factor leading to the complexities of the denominator).

Members of both classes of species produce recombinant offspring by randomly choosing a parent for each gene (in this case, a gene is a single double precision floating point

number found in A, V or T). These are used to produce a new organism x, along with a new target mate appearance - its own colour, $c_x(x)$ - it is assigned a new unique ID. The inherited values also gain mutations in the form of the addition of a normally distributed random number with standard deviation σ_{SV}, σ_{SA} or σ_{DV}, σ_{DA}, depending on the class and variable type. Resultant values outside of the interval $[0, 1]$ are brought back into range by applying either $f : x > 1 \to 1, x < 0 \to 0, x \in [0, 1] \to x$ or $f : x \to x \mod 1$ [1]. Mutations are applied with frequencies m_D and m_S to the drifters and signallers respectively.

E.1.1 Selection Rules

The species in each class have different selection rules, the 'drifters' $\mathcal{D}$ merely drift, their visual pigments and target colours are largely irrelevant. The populations of each drifter species are modified at each generation by applying the following rules a given number (r_D) of times:

```
1: Select a random member of the population for a drifter species
2: Select a random mate from the same population
3: Recombine their genes and mutate
4: Replace a random member of the population with the new offspring
```

The 'signallers', $\mathcal{S}$ are evolved by rules chosen to mimic sexual selection. Instead of knowing the species to mate with, they are required to judge it by their appearance from all of $\mathcal{S} \cup \mathcal{D}$. To do this they use a psychometric function $\tanh(kd)$, where k is a constant and d is the euclidean distance between the colour (c) of their potential mate and their target mate (T). This function fits into the following evolutionary schema, which is applied to r_S organisms each time step:

```
1: Select a random member from the population of a signaller species
2: Choose a potential mate from any species
3: Find the value of the psychometric function
4: Choose a uniform random variable from [0,1]
5: If the random variable is greater than value from the
    psychometric function, go back to 2
6: If the mate is not of the same species, do nothing, otherwise
```

[1] mod 1 is the floating point modulo operation in base 1, setting the value before the decimal place to zero.

```
7: Recombine genes and mutate
8: Replace a random member of the population with the new offspring
```

E.1.2 Establishing Convergence

The convergence of the simulation was judged by means of a perceiver independent proxy for colourfulness, H. This is fast to calculate and allows a simulations progress to be tracked. It is the sum of $a \log a$ over all elements of the appearance for each organism in each species:

$$H = \sum_{s \in \mathcal{S} \cup \mathcal{D}} \sum_{x \in s} \sum_{a \in A_x} a_i \log a_i \tag{E.1.4}$$

The fixation of this value to within an interval the size of 5% of its deviation from the start, over a period of 100 generations, was taken to be sufficient for the simulation to be ended. Though generally simulations continued beyond this point.

E.1.3 Analysis

The appearance of the species was summarised in terms of colourfulness (see chapter 2), the fractional distance of a point from the centre to the edge of the colour space. The frequency of colourfulness was summarised by taking the colourfulness of each organism within $\mathcal{S}$ (or $\mathcal{D}$) as judged by every organism in $\mathcal{S}$ (or $\mathcal{D}$), producing n_S^2 (or n_D^2) values which are then used to create histograms. The histograms are then corrected so that they would be uniform were the organisms uniformly distributed across the colour space, the histogram thus represents the organisms *density on the colour space*. To perform this correction we note that the locus of equal colourfulness is the boundary of the colour space scaled by the colourfulness, the volume (area) with colourfulness less than a c is proportional to c^{n_e}, in the case of the dichromatic species in this simulation $n_e = 2$. So the volume of colour space corresponding to a histogram bin corresponding to the colourfulness interval $[c, c + \Delta c]$ is $k\Delta c(c + \Delta c)$, where k is equal to the total volume (area) of the colour space. The normalised histogram is invariant with respect to the value of k, and determination of its value is not required. Because of this we only need to multiply each bin by $c + \Delta c$.

E.2 Behaviour of the Model

The first thing to consider in the model is the behaviour of the species that are not under selective pressure for appearance. The lack of pressure on appearance causes the species to distribute uniformly across the space of spectra. When these are projected to form colours

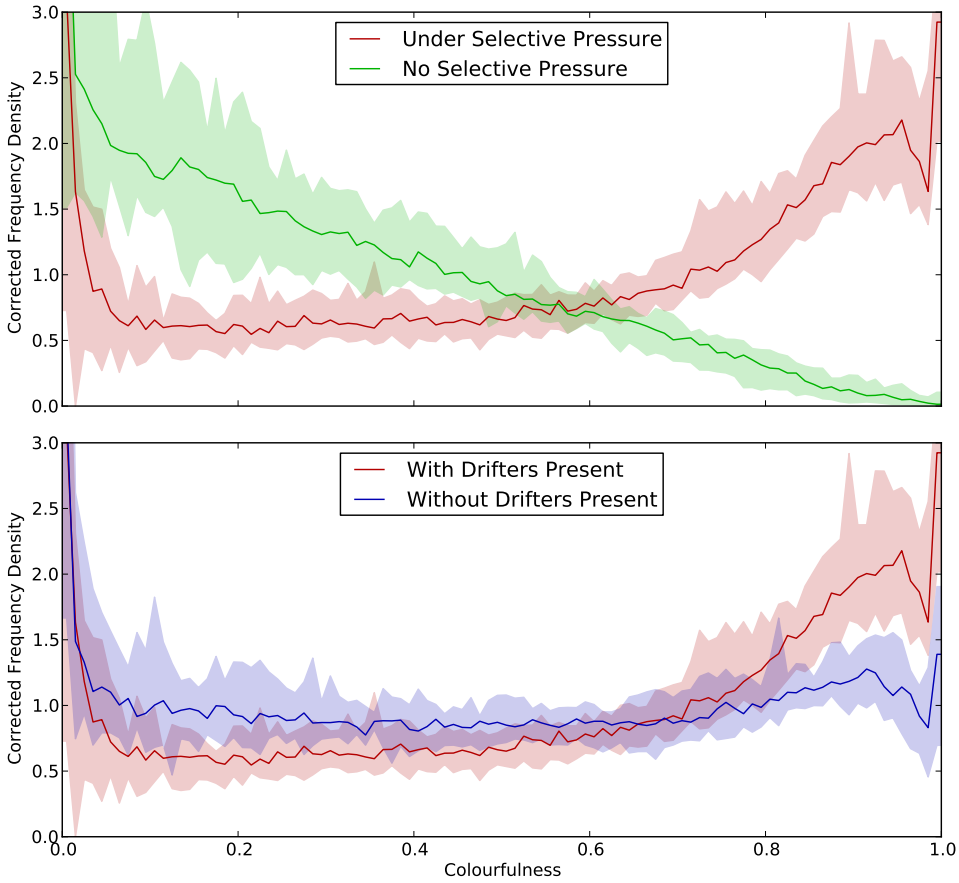

Figure E.1: Data from the sexual simulation model with 42 species (plus 42 drifting species when present). These plots show the frequency that a part of the colour space with colourfulness of a particular value is contains a member of a species of a given type. The data is corrected for the increased sampling of colourfulness values at higher colourfullnesses (e.g. there is more area for a colourfulness of 0.1 than 0.9). In both graphs the lines represent median values for ten simulations and the filled area indicates the full range of the simulated results. The red lines show the colourfulness under selective pressure from drifting species; green, the colourfulness of the drifting species themselves; and blue is the sexually selected species but without the presence of the drifters. The simulation was performed with a psychometric constant of 4, spectral dimensionality of 4, 2 visual pigments, standard deviations of Gaussian mutation sizes of 0.1. The probabilities of mutation is 0.05 for spectra, 0 for visual pigments (these generally fixate within 100 generations to a pair of fairly orthogonal pigments). There are 30 organisms per species and 5 are potentially replaced per generation.

we see that more of them project to the centre of the space. This then presents itself as a net movement towards the centre of the colour space. This can be seen in figure E.1 as the high probability density for low colourfullnesses in the green curve. It is represented in figure E.2 parts *a* and *b*.

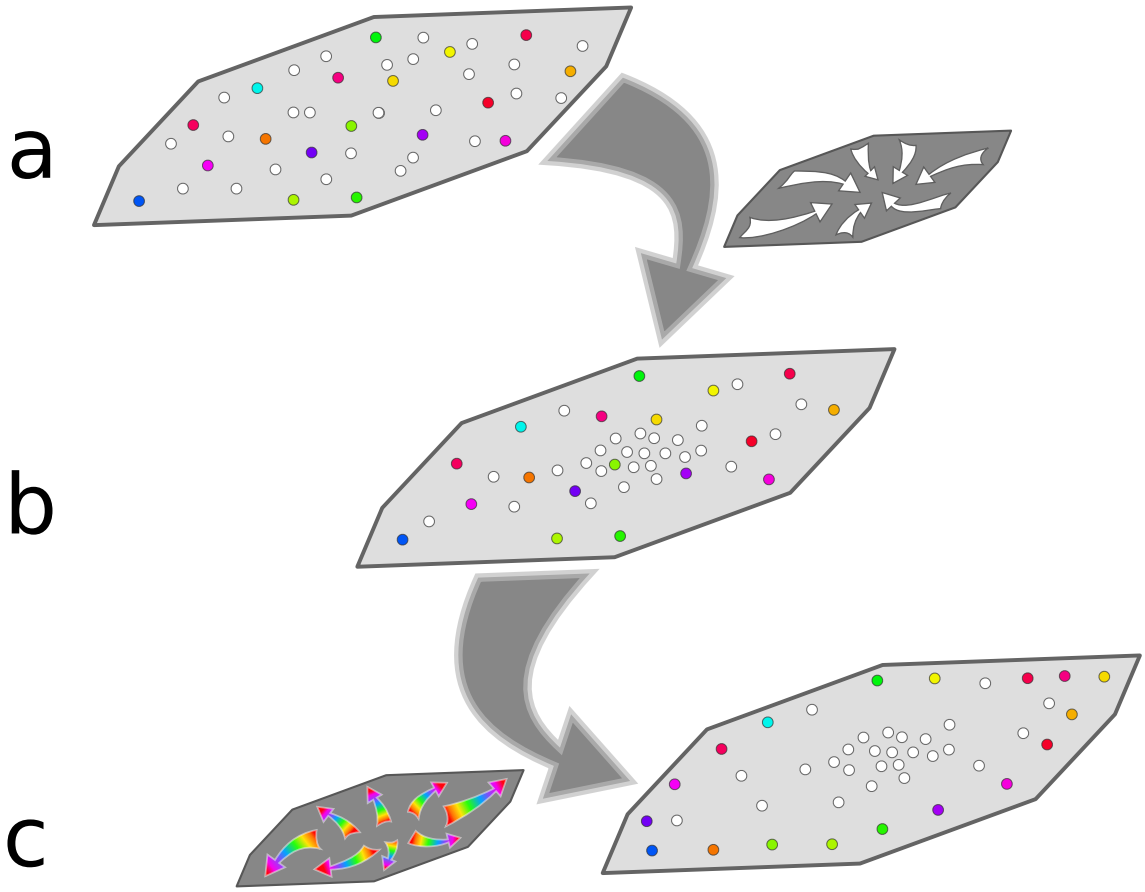

Figure E.2: A caricatured representation of the causal mechanism of strong colour formation in the sexual selection model. Multicoloured dots represent the species under sexual selection, i.e. those required to identify their conspecifics using colour cues. The white dots represent species with no selective pressure upon their appearance. (a) The species begin at random points in colour space. (b) The species that are not under selective pressure for colour show a net drift towards the centre of the colour space. (c) in response to the increased number of competing signals in the centre of the space the species required to communicate using colour move towards the outskirts of the colour space. In reality the two processes happen near simultaneously, the arrows represent the causal relationship only.

The response of the organisms under sexual selection is a net movement away from high species density (see red curves in figure E.1), which is towards the colourful edges of the colour space. There are a couple of other features that can be seen in the data.

There is a slight tendency towards the edges of the colour space due to the bias induced from truncating the reflectance genes when a mutation takes it outside of the range $[0, 1]$, this can be seen in the blue curve in figure E.1. In addition there is also a bump induced from the regularity of what is effectively an approximate packing of species from the edge inwards. Some species move right to up to the edge of the space, then there is an area of lower density. Following this there is then a higher density again. This is just a manifestation of requiring a particular separation between species and aligning them with the edge of the space.

This behaviour happens across a range of parameters. Whilst it is clearly unreasonable to expect to find a real word situation where there are a large number of species only differing by their colour, it is useful to consider such a system for the purposes of demonstrating the selective pressure from the the non-selected organisms. When the physiology responsible for colour formation is not closely modelled (like it is in chapter 7) and a direct Gaussian mutation scheme with truncated boundaries is used, there is a tendency for species to 'stick' to the edge. The probability of being exactly at the edge is greater than at any other position[2] biasing the results with magnitude negatively correlated with species number. The addition of extra sexually selected species allows this effect to be diluted to the point of making reasonable comparisons, and in addition, easing the production of good quality histograms.

E.3 Algorithmic Calculation of the Object Colour Solid

The following algorithm is a 'fast' way of performing the accurate calculation of the colour solid without any constraint relating to the convexity of the spectral line (unlike the extreme spectrum method). The direct calculation of the convex hull of the spectrum space is intractable for highly resolved spectra, as the number of extreme points to consider $\#\mathfrak{T}$ grows exponentially with the number of samples of the spectrum, m:

$$\#\mathfrak{T} = 2^m \tag{E.3.1}$$

So, even for a 2D colour solid being calculated on a PC with 4GB of free memory, representing each point as a pair of double precision floating point numbers, we are limited

[2] All the positions beyond the edge map exactly to the edge. i.e. the probability moving onto the edge (of one spectral dimension) from position x is $\mathrm{erf}(-|x-1|/\sigma) + \mathrm{erf}(-|x|/\sigma)$, all other positions have infinitesimal probabilities.

to 28 samples, simply by what we can represent in memory.[3]. This says nothing about how long it will take to run; there is no convex hull algorithm with complexity less than $n \log n$, so this naïve method is clearly at least[4] exponential time in number of samples.

Often, this kind of problem is solved using quadratic programming methods (Ohta and Wyszecki, 1975; Wyszecki and Stiles, 2000), but here I propose a different solution, whose accuracy is limited by how accurately we represent the photoreceptor sensitivities, not by how many points we calculate.

It works by calculating the projection of the reflectance spectrum hypercube directly, at every step increasing the number of points twofold. However, at each step a convex hull algorithm is used to prune as many points as possible, preventing what would otherwise be a combinatorial explosion. An optimisation is also possible at line 6: in addition to the black point (the origin), the starting set can be made to include the white point ($\mathbb{1}$). This significantly enhances the efficiency of the pruning.

```
 1 function solidCalculator(photoreceptor_response_vectors)
 2   # Create a matrix of photoreceptor responses
 3   let columns of response_matrix = photoreceptor_response_vectors
 4
 5   # Create a starting hypercube
 6   let point_list = [point at origin]
 7   let i = 1
 8   while i <= number of photoreceptors
 9     translated_points = point_list + ith row of response_matrix
10     point_list = point_list union translated_points
11     increment i
12
13   # Iteratively apply convex hull algorithm
14   while i <= number of elements in response vector
15     translated_points = point_list + ith row of response_matrix
16     point_list = point_list union translated_points
17     point_list = convex_hull(point_list)
18     increment i
```

[3] A pair of double precision floating point numbers requires $16 = 2^4$ bytes of memory. $4GB = 2^{32}$ bytes. So solve $2^m \cdot 2^4 = 2^{m+4} = 2^{32}$.

[4] A rough calculation suggests that it is greater than exponential time, $\lim_{n\to\infty} \frac{n2^n}{2^n} = \infty$, but slower than factorial time $\lim_{n\to\infty} \frac{n2^n}{n!} = 0$.

```
19
20    return point_list
```

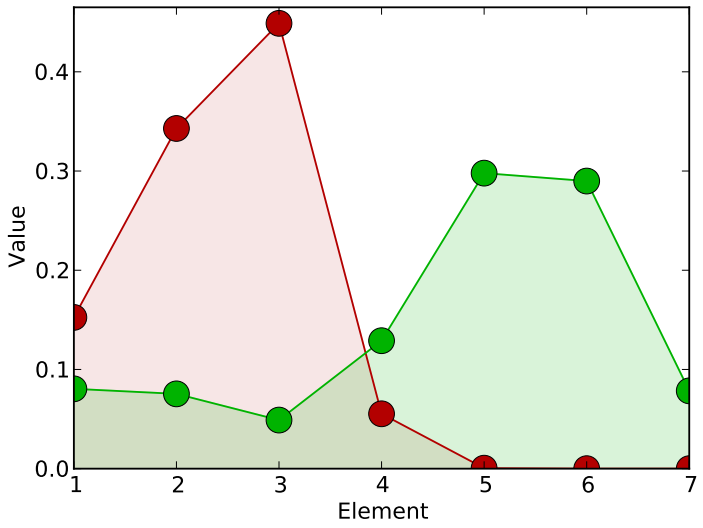

Figure E.3: Data based on A1 pigment templates (Stavenga et al., 1993) used in the calculation in figure E.4. Each curve represents a vector of seven elements with values of each element given by the y-axis.

How can we be sure this works? To assure this algorithm is correct we need only to argue that it will necessarily visit every point on the hull, as the convex hull algorithm can never remove a point that is actually on the final hull. Let us denote the solid represented in the calculation at step i by s_i, and the value of the photoreceptor sensitivities at step i by v_i. With some calculation[5] it can be shown that:

$$s_{i+1} = \mathcal{C}\left(s_i \cup \{x + v_i \ : \ x \in s_i\}\right) = \mathcal{C}\Big(\mathcal{H}\Big(\mathcal{H}(s_i) \cup \{x + v_i \ : \ x \in \mathcal{H}(s_i)\}\Big)\Big) \quad \text{(E.3.2)}$$

implying that we do not 'loose' any points by iterating using the hulls, $\mathcal{H}(s_i)$, rather than the full set s_i. So at each stage of the iteration, even though we are removing points that were on the hull in previous steps, we are still keeping a full representation of the full colour solid up to that point.

[5]This can be achieved by writing the convex coverings and hulls explicitly in their summation form.

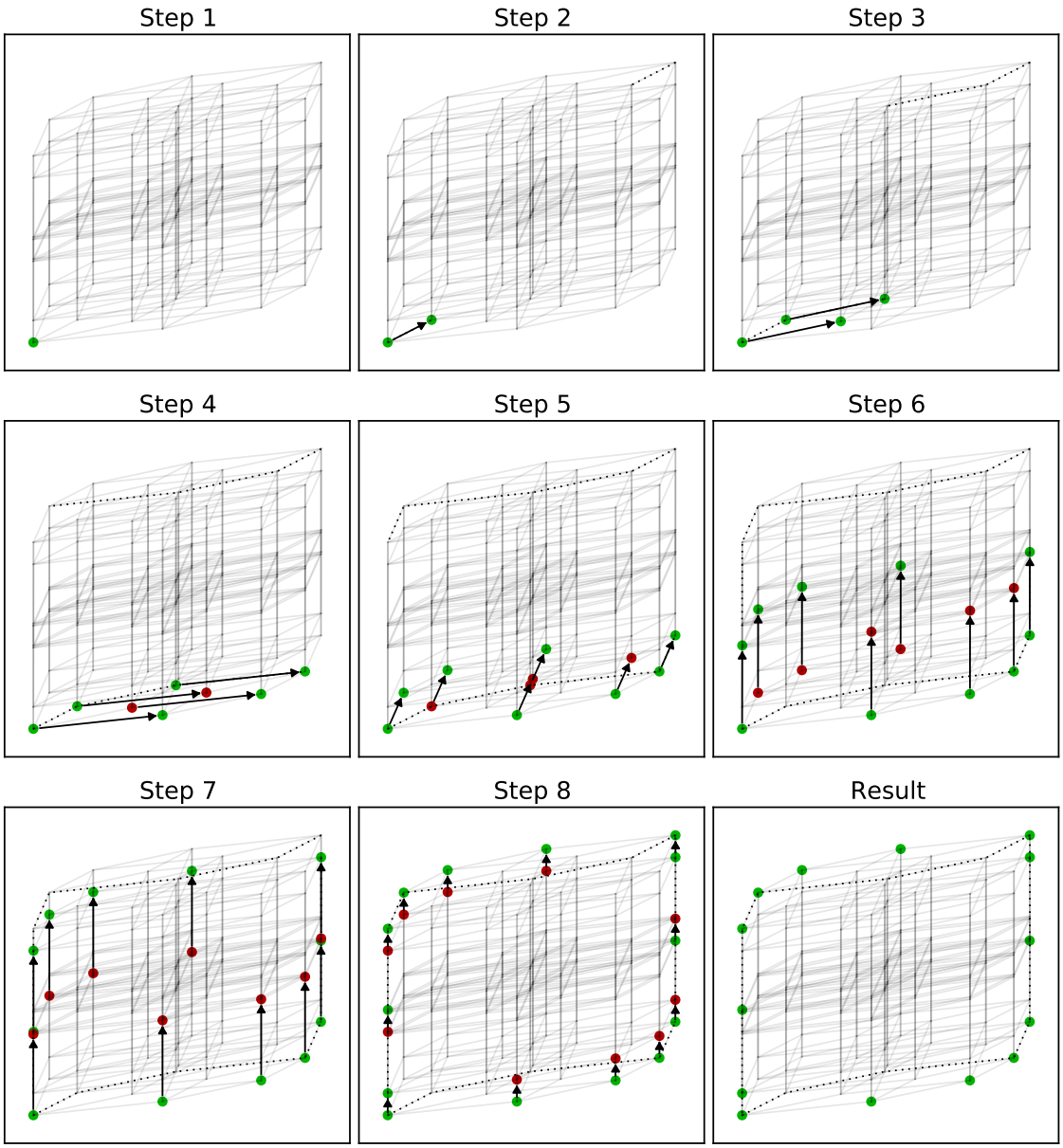

Figure E.4: Calculating a 2D colour solid. We begin with a single point (Step 1) then create a projected m-cube (a square) by successively duplicating the number of points by making a copy and adding a vector (arrows). After the shape has the same dimensionally as the space it lies in, we begin pruning points after each duplication (Steps 4-8 inclusive). Points that will be removed are shown in red. This makes the algorithm possible. The dotted lines show the coordinates of the 'extreme spectra'. We can see that one branch of algorithm follows the same path as these, however, this path does not lie on the convex hull, so some of the points on it get pruned away (all of them that do are removed are red in step 5). The light black lines are all the edges of the hypercube, most of which do not get traversed due to the pruning.

Appendix F

More Renderings of the Phenotypic Space

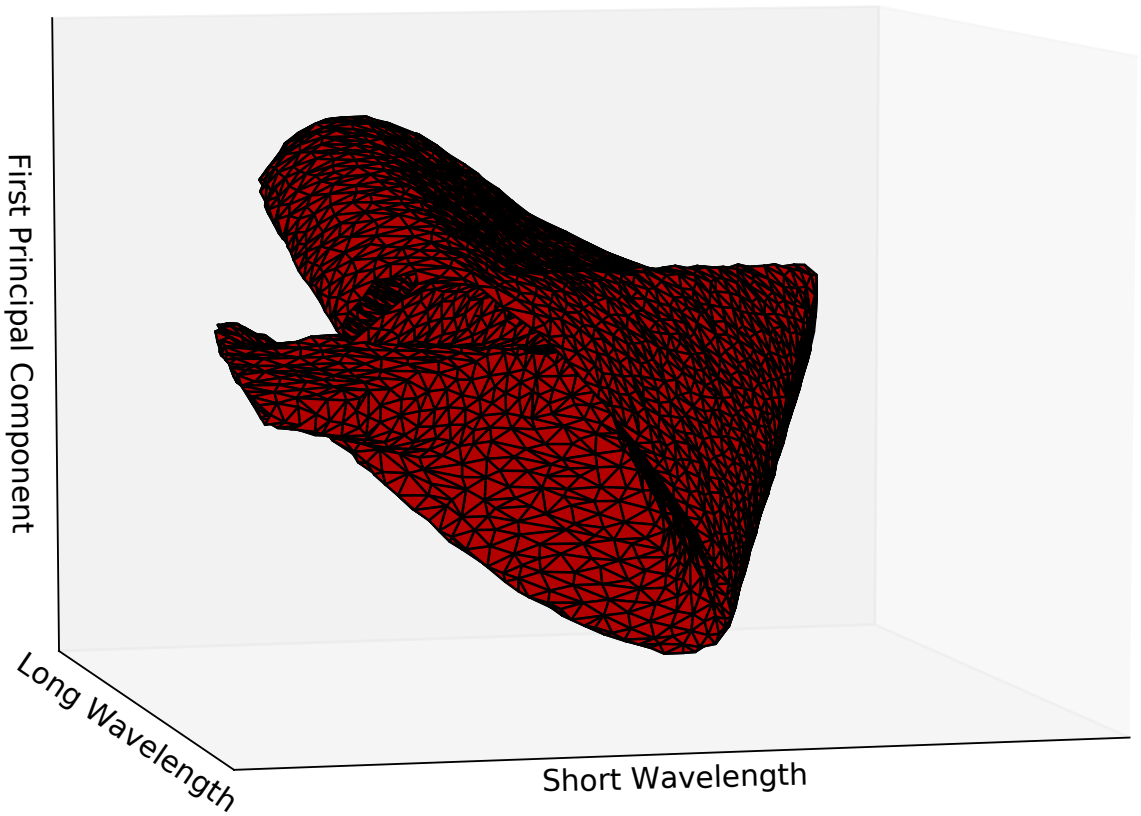

Figure F.1: Rendering showing some of the complexities of the relationship between the physical parameters, the colour, and the hidden space.

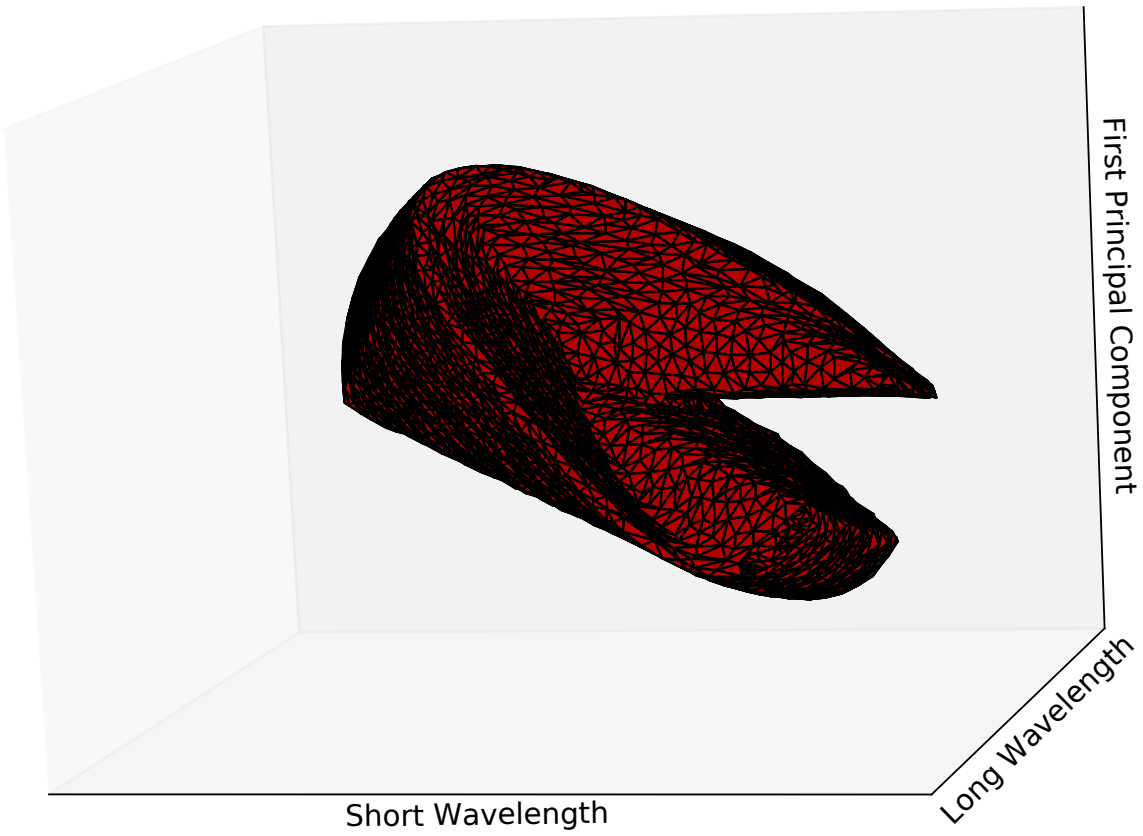

Figure F.2: A rendering showing the two different zones that can achieve the same colour.

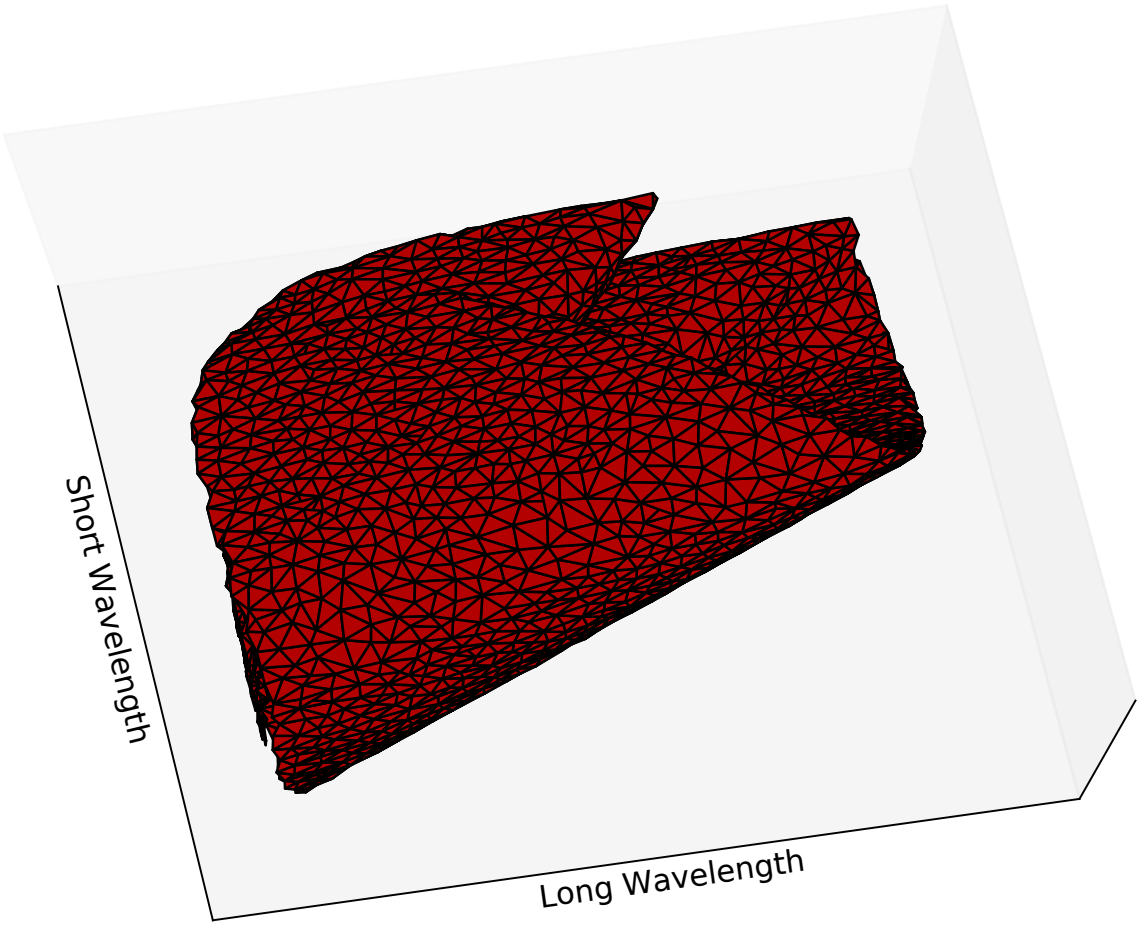

Figure F.3: Another rendering showing how non-linear the space is.